POLITECNICO DI TORINO

Master Degree in Mathematical Engineering

Master Degree Thesis

The demand for public transport: analysis of mobility patterns and bus stops

Supervisors

Prof. SILVIA CHIUSANO

MSc ELENA DARAIO

Candidate

ANDREA ATTILI

Academic year 2020/2021

Abstract

The availability of smart card data from public transport allows analysing current and predicting future public transport usage. As an essential part of public transportation, forecasting bus passenger demand plays an important role in resource allocation, network planning, and frequency setting. In this thesis, demand and offer of a public transport consortium located in north-western Italy have been taken into account. Also, demographical data about the customers and the land, and historical weather conditions of the area during the period of interest have been collected. Two main kinds of analyses of the demand have been performed: at first with respect to many possible features, such as type of day, hour of day, stop point, trip, kind of user (particularly in terms of age), kind of travel document. This has shown that three main kinds of days can be identified: working, half-holidays (like Saturday) and holidays; moreover, the demand shows a great variability across different trips, stop points and users and there are two daily time-slots with a peak of validations. Among the customers, the majority is represented by students. Then, the focus has been moved to the analysis of the stop points, which have been grouped according to the incoming demand, through a clustering algorithm. The proposed methodology has been developed in Python with the support of QGis software for geographical visualizations. The results may be the starting point for an efficient forecast of the demand, at any stop point and time interval. The implications of this thesis could be appealing for public transport operators, since forecasting the passenger demand is necessary to properly plan on-demand mobility services, increasingly being promoted as an influential strategy to address urban transport challenges in large and fast growing cities.

Acknowledgements

For the precious support during the whole work, thanks to Links Foundation, represented by Maurizio Arnone, Brunella Caroleo and Alexander Fazari.

For having provided data and related explanations, thanks to GrandaBus, represented by Mauro Paoletti and Ivo Rinaudo.

Table of Contents

List of Figures

Chapter 1

Introduction

Estimating the public transport demand has become a great concern for the agencies which offer this service: it allows to improve the offer and, as a final goal, to optimize the economical resources. However, it is a very difficult job, since the number of passenger which need to travel at specific place and time depends on several factors and therefore it is subject to a great variability. The forecast is particularly useful just when this variability is higher, because average-based methods fail, so some more complex techniques, related to machine learning and artificial intelligence, are needed. This happens at peak hours, when the transport service must be properly intensified: if the offer is too low, the transport means will become overcrowded and it will probably delay, causing inconveniences to the passengers (both on board and waiting at the stop points), that, as a consequence, will switch to other solutions for their journeys. If instead there are more trips than needed, the agency will probably lose money, because some of the trips could be cleverly avoided. Of course, an economical lost is also an indirect consequence of the previous scenario, so at the end finding a good compromise is crucial for the agency.

A good solution, which may be enhanced by the ability of predicting the demand in real time, is a service on-demand, where trips are properly intensified when really needed, with criteria to be defined. Nowadays, on-demand transportation mainly involves car or bike sharing or other similar services (while experiments related to on-demand public transport are still limited), which involve one or few people at a time: usually, they look for the closest free vehicle, take it and leave it in the most suitable place with respect to their destination. The same service with a bus or another public means is certainly more complex to set up: in order to fully exploit the capacity of the vehicle, the needs of dozens of people must be put together. Their travels should be similar in terms of space and time, so it could be useful also to group passengers based on the spatial-temporal coordinates of the journey, but also on its reason: unless they are individually interviewed, this can be just

inferred from the regularity of the journey across the days (provided that one will always use the same travel document) and, if available, from their demographical information (especially age and possibly some clues about his occupation). For example, some students living close one to the other and going to the same school, probably at the same time, would take benefit from this solution.

From this perspective, the task is more challenging for isolated places and at hours with a lower demand. Actually, in this scenario the stop points are farther from each other and there is a smaller number of people moving. Therefore, less trips are necessary, so they must be organised with more precision, because passengers may be discouraged from using a public service if there is too much distance, either in time or space, between his needs and the effective planning of a certain trip.

In order to estimate the demand for a certain trip, smart cards play a crucial role: however, this is not the only possible travel document. Specifically, for occasional customers it is more convenient travelling with single tickets, which usually don't expire, and these are often physical tickets, not associated to a smart card, so the validation is recorded, but it doesn't have references to the customer. On the other hand, nowadays smart card are spreading for charging subscriptions and transport credit documents: the first refers to a fare which gives the right to travel, within a certain area, for a certain time of validity, which can start with the purchase or with the first validation; in the latter case, a certain amount of money is charged on the card and, at each travel, depending on the departure and the arrival place, the corresponding price is subtracted.

As giving more information about users and their habits, passengers should be enhanced to use smart cards, rather than physical tickets, possibly with favourable fares. The smart cards are usually validated only when boarding on a public means, especially in case of subscriptions. This doesn't allow to know with certainty the final destinations of a journey, which however can be inferred from the following validations of the same person. Since also the timestamp is recorded, it should be easy to understand if two consecutive validations belong to the same journey (so in the meantime there was just a transfer), or not: in the latter case, it is expected a symmetry between the journeys of the same day and, depending on the kind of passenger, also a regularity across different days. Information coming from the smart cards are therefore precious, if properly exploited: with the purpose of forecasting the demand, validations for a long period of time should be collected, together with all the possible variables which can affect their trend. However, it is possible that at some stop points the demand is quite stable, so the prediction can be done also with baseline methods, where no other indicators (apart from the number of validations) is considered. These may consist in predicting the future demand simply by using the moving average of the most recent observations at the same stop point, or, slightly more sophisticated, applying the moving average

on the noise, which, together with the trend and the seasonality, results from the decomposition of the time series of the demand. On the other hand, wherever the demand shows an unpredictable trend along time, a more sophisticated analysis should be conducted and the relationship between the dependent variable and each possible predictor must be carefully studied.

From what said above, it sounds that the prediction should be preceded by two kinds of analysis:

- Observing how the demand varies if one or more of the candidate predictors are changed: at this step, the stop point is just one of these variables
- Based on the single stop point, its features must be carefully understood. For this purpose, all the other candidate variables are discarded and the demand is observed across different temporal dimensions, in order to put together the stops with a similar behaviour (with the hope that the same prediction algorithm will be suitable for all of them) and to find where some more sophisticated techniques will be needed

The remainder of this thesis is organised as follows: Chapter 2 gives a general review on the main problems related to improving the public transport. In Chapter 3, the followed methodology, the softwares used and all the available data are described. In Chapter 4, there is a necessary preliminary operation consisting in cleaning the data. In Chapter 5, the first step of the analysis introduced at the previous paragraph is performed; Chapter 6 describes the second step of the analysis, while Chapter 7 contains a discussion of the results obtained and finally in Chapter 8 there are some conclusions and suggestions about how to best exploit this work.

Chapter 2

Review of the literature

Organising public transport is a complex task, which in turns includes a lot of subproblems. Actually, forecasting the demand is not the only possible strategy for an improvement. Here is a list of the most relevant open issues, each with references to some contributions, which makes it clear that they all have their own importance:

1. **Segmenting customers** based on the regularity (both spatial and temporal) of their trips: see for example [**10**], where the partitioning is preparatory for solving also problem 4, and [**25**].

2. **Forecasting mobility at a certain stop, station or zone**, which means estimating how many people will need a certain mobility service, at a certain place, during a certain timeslot (prediction for a time instant is very difficult): further specifications are given below.

3. **Estimating the most likely destination** (which means the alighting stop) of a passenger of the public transport, given information about the boarding stop, but also about the starting point of his following trip during the same day, if present. An answer to this question, which can be useful for the previous problem, is suggested in [**26**], [**27**], which give also a description of the virtual-checkout algorithm mentioned in the following and of its theoretical basis, but also in [**29**] and [**30**]. The latter introduces a methodology useful also for unlinked trips, which means that the origin of the following trip of the same day is unknown or it is not done at all, and [**31**], where a neural network is used for the prediction.

4. **Forecasting individual mobility**, which means trying to guess, for each person, in a certain day, whether he will move and, if so, when he will do it and his origin and destination. A useful tool to represent the aggregated result

of this analysis is given by the origin-destination matrices, which, in their most basic form, have a list of origins on the rows and a list of destinations on the columns. The value in position (i,j) corresponds to the number of people moving from i to j during a certain time interval. It is possible to extend this methodology, in order to include more information: for this purpose, more than two dimensions are needed, thus turning to tensors. In this case, hidden relationships among different variables can be found by decomposing the tensor. A mathematical explanation of tensors and their factorisation is given in [**23**] and [**24**], while some possible applications are shown in [**1**] and [**2**], where also the correlation between the variables is taken into account. Instead, [**6**] and [**13**] deal with this problem by using a neural network with long-short term memory, while [**10**] tries to estimate the destination of each previously formed group of customers. Finally, in [**15**] there is an attempt to recover the chain of journeys of each one by applying Markov models.

5. **Estimating the time instant at which a certain vehicle of a certain route will get at a certain stop**, based on the past travel times. Clearly, this problem is strictly correlated with (5); possible solutions are described in [**11**] and [**14**]. The first chooses to implement a variant of the SVM algorithm, the latter is based on clustering the historical data.

6. **Predicting travel and dwell time**, which are the times needed for a public means to travel between two consecutive stops and to let passengers boarding and alighting at a certain stop, respectively. Further explanations can be found in [**8**], where the Least Squares SVM is used for the prediction, which is preparatory for solving the next problem; in [**16**] a neural net takes as inputs for this purpose the estimation of the traffic and of the incoming demand, while in [**17**] the travel and dwell times are clustered based on their distribution along the routes.

7. **Predicting bunching and preventing it**. Bunching occurs when two vehicles running on the same routes simultaneously arrive at the same stop: this may cause the overcrowding of the first vehicle and an increased headway (and therefore an unexpectedly great waiting time for people at that stop). This problem can be addressed together with the previous one, so [**8**] is again useful.

8. **Reorganising the routes of the public transport**, based on travel and dwell times and on demand at each stop, as done in [**21**] for Baghdad (Iraq), where origin-destination matrices are again used to represent the journeys.

Problems 1-4 deal with the demand of mobility, which involves not only public transportation, but also some private or sharing services; demand is important as

the offer (which is the macro-topic of problems 5-8), since at the end both categories of problems have the same goal, that is optimizing the public transportation service.

As already said, this thesis provides the starting point for the second problem, which in turn can be seen from different perspectives. For example, in [**4**] the goal is predicting the demand of a car-sharing service in Turin; in [**5**] the demand forecast is restricted to special events, which attract a lot of people, so such a demand is very variable; in [**9**] there is an attempt to predict future position of taxis in Macao; in [**12**] mobility is analyzed on a regional scale in Beijing. However, most of the previous works ([**3**], [**7**], [**18**], [**19**], [**20**] and [**22**]) is concerned with predicting the number of people waiting at a certain stop for a certain public transportation route: [**3**] works with time series and describes in details the variables and the hypothesis of the problem; [**7**] is oriented towards short term prediction by using a Poisson process based on a neural network; [**18**] analyses the correlation between the target and the possible predictors before performing the forecast (for which a neural net is again used); [**19**] insists on some unusual features of the stops, while the forecast occurs through a multiple regression; finally, in [**22**] there's a comparison of the results obtained with several methods, mainly related to trees and regression.

Finally, [**32**] and [**33**] are very recent works where the impact of Covid-19 on the use of public transport is analysed: specifically, the first depicts some new mobility patterns which may be retained also after the pandemic, in the latter the population is clustered according to how much its travel habits have changed from before to during and after the pandemic.

However, none of these works has analyzed the trend of the demand with respect to all of the possible variables which can affect it (or, at least, they don't describe it explicitly). Also the clustering of the stop points has been performed in another way, mainly due to the land features: in other documents the setting consisted in one or more big cities, while in this case there are several small or medium towns, with the exception of Turin, which, however, is in a marginal position, not only geographically, but also in terms of total demand.

Tables 2.1, 2.2, 2.3 respectively synthesize the tackled problems (together with the geographical setting), the applied techniques and the used variables in the majority of the papers taken as reference. Some documents are excluded from the tables because they have other aims. From the first two tables, it seems that the most recurrent techniques when trying to predict the demand of public transport are those related to neural networks, time series and regression, while clustering is mainly used for segmenting customers or stop points (the latter has been included as part of problem 2). In the last table, the variables are also divided according to the type: offer (in blue), demand (in green), weather (in yellow).

	DEMAND PROBLEMS				OFFER PROBLEMS				
Doc.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Location**
[1]				■					Singapore
[2]				■					Shenzhen
[3]		■							Yantai
[4]		■							Torino
[5]		■							Nanjing
[6]				■					Rennes
[7]		■							Australian city
[8]						■	■		Beijing
[9]		■							Macao
[10]	■			■					Porto
[11]					■				Beijing
[12]		■							Beijing
[13]				■					Japanese city
[14]					■				Nashville
[15]				■					London
[16]						■			Singapore
[17]						■			Gran Canaria
[18]		■							Tokyo
[19]		■							Madrid
[20]		■							Madrid
[21]								■	Baghdad
[22]		■							Buenos Aires
[25]	■								Gatineau, Canada
[26]			■						Cuneo's province
[27]			■						Cuneo's province
[29]			■						Gatineau, Canada
[30]			■						Gatineau, Canada
[31]			■						Seul

Figure 2.1: Topic and setting of each of the reference documents.

MODELS	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[19]	[20]	[21]	[22]	[25]	[26]	[27]	[29]	[30]	[31]
Neural nets						■	■	■	■		■	■	■			■		■										■
Clustering										■		■		■			■						■					
Random forest				■				■		■												■						
Tree										■												■						
Bagging																						■						
Grad.boosting				■																		■						
KNN								■			■																	
SVM and var.								■			■																	
Linear/log regr.				■											■				■	■		■						
Time series			■	■	■						■																	
Poisson model							■									■												
N. Bayes										■																		
Top K																												
Corr. Analysis		■																■		■								
Tensors	■	■										■									■							
Markov									■						■													
Gaussian regr.								■															■					

Figure 2.2: Techniques used in each of the reference documents.

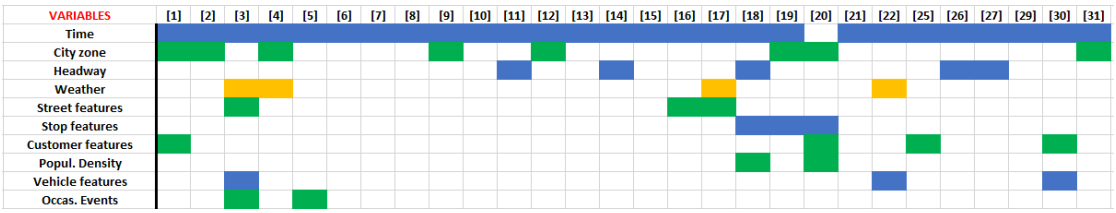

Figure 2.3: Variables used in each of the reference documents.

Chapter 3

Materials and methods

3.1 Methodology

Two main programmes have been of support for this analysis. Python has been used for the various computations, together with several useful packages, such as pandas for managing tables, numpy for arrays and matrices, csv for reading and writing csv files, gtfs_kit for analysing the GTFS tables, time and datetime for working with temporal data, geopandas and geopy for retrieving some geographical information, re and requests for downloading data from web pages, sklearn for applying and evaluating the clustering algorithms, matplotlib for plotting the graphs. Moreover, OpenCage Geocode is an API which converts any place in the world to its coordinates and viceversa, while the software QGis has been used for the geographical representation of some results. Actually, starting from a file showing the position of each stop point and another, used as background, with the geography of the roads of the involved land, it has been possible to change the color of the localities based on some attributes, or to do the same with the stop points, just by merging different tables.

The followed methodology is detailed on the left of Figure 3.1, together with the related Chapter and the main steps of each part.

3.2 Data description

This thesis mainly focuses on public transport data, dated back to 2019 and provided by Granda Bus, a consortium of 16 transport agencies, mostly operating in the area of Cuneo, in North-Western Italy, but it partially covers also five neighbour provinces: Turin, Asti, Alessandria, Imperia and Savona. From the offer side, the GTFS (General Transit Feed Specification) format is adopted. This is a commonly used representation of the planned timetable of public transport services,

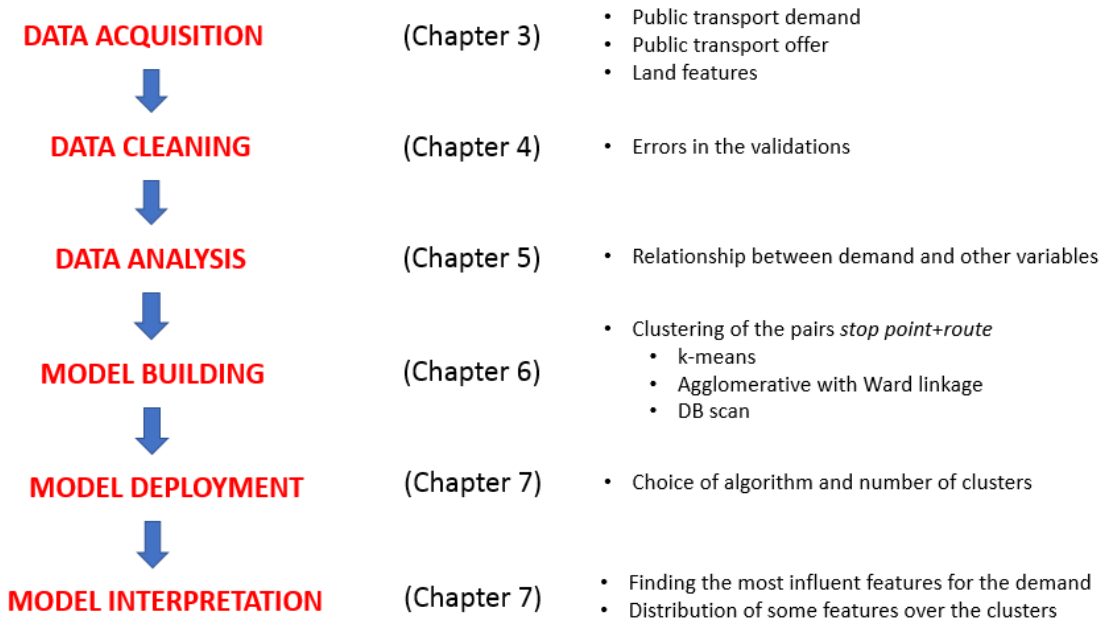

Figure 3.1: Workflow of the methodology followed in the thesis.

also with geographical information. The service includes 237 routes, 6069 trips and 7371 stop points. In this case, the focus is on a sample month: as not including particular holidays, October has been chosen.

For the year of interest, data have been collected through personal smart-cards and they involve passengers, their travels with public transport means and the travel documents used. For each validation, timestamp and stop point are recorded. In this way, many of the variables reported in Figure 2.3 can be in turn collected. Specifically, for what concerns stop points, apart from what explicitly given (position of each stop point), some other useful information can be recovered, including the position in the trip, possibility of interchanges, weather condition referred to the moment of the validation, population density of the surrounding land. Also for what concerns the customers, data coming from different sources have been integrated. Actually, for those associated to a travel document, some useful demographical information are explicitly given, while other ones can be easily inferred by the data (in particular, from the fiscal code). Unfortunately, not every smart-card is related to a person (specifically, the codes starting with character "2" are impersonal and don't appear in the users table); viceversa, each customer may hold more than one travel document. The variables not considered with respect to Figure 2.3 (mainly because not available) are the conditions of the roads and of the vehicle and indications about special events in the neighbourhood.

The recorded information include nearly 1000 travel documents and 100000

users. However, less than one third of them have travelled in the period, apart from about 15000 people with an impersonal document.

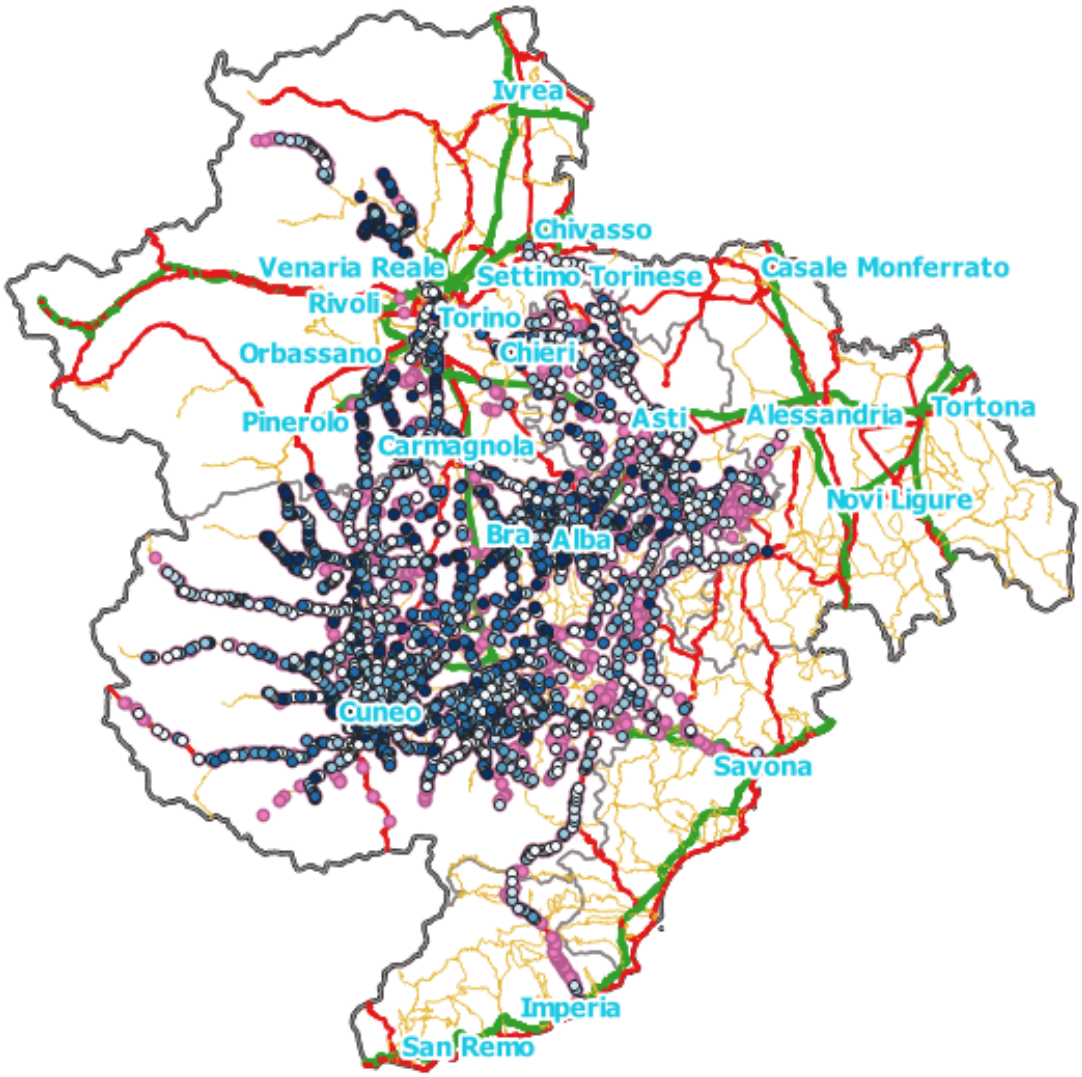

Figure 3.2: Geographical distribution of the stop points.

As shown in the map (Figure 3.2), where the color of the stop points ranges from white to blue depending on the number of incoming validations (at least one), while stops which are never associated to a validation are in pink, travels are concentrated in Cuneo; a smaller number occurs in Saluzzo (not shown in the map but located in the halfway between Pinerolo and Cuneo), Torino and Alba. There is also a partitioning into the six provinces (boundaries are the grey lines) and a representation of the main roads, whose importance depends on the

thickness and the color of the line. These have been downloaded from [**41**], where all the roads of North-Western Italy are present: therefore, those outside of the boundaries or of low importance were excluded: specifically, those of kind *motorway*, *motorway_link* (both green), *secondary*, *secondary_link* (both red), *primary* or *primary_link* (yellow) have been retained. Moreover, the cyan labels allow to identify the main towns.

The available data can be grouped as follows:

- **Public transport demand**. Smart-card data provide information about the timestamp and the location of the stop for each boarding passenger. For travellers with transport credit, the same information are available also for the alighting stop; in the other cases, they are all estimated through the virtual check-out algorithm described in [**26**] and [**27**]. Data are organised into five tables:
  - *validazioni* collects, whenever a passenger boards on a bus or alights with a travel document, the timestamp and, if available, the code of the travel document, of the bus stop, vehicle, route and trip. Moreover, the type of validation (see *codici_validazione*), the number of people (passengers with transport credit may simultaneously validate for others, apart from themselves) and the code of the passenger are available.
  - *biglietti* has the same attributes of *validazioni*, but it involves only tickets bought on the bus and in this case the code of the passenger is not available. Moreover, in this table the stop point is not precisely recorded, since the price of the ticket depends on the location where it has been bought, not on the exact starting point.
  - *titoli* (sometimes called *documents*) contains, for each travel document, a short description, the type (single ticket, carnet, subscription), whether it is specific for a certain category of people (students, retired, people with disability and their carers) and how long it can be used.
  - *utenti* (sometimes called *users*) contains demographic data about the passengers, including also past users or very young ones (maybe travelling on school buses), each identified by the same code used in *validazioni*: the town where he lives, the related postal code, gender, age (which, due to what stated above, ranges from 0 up to beyond 100) and some characters of the fiscal code (from the 12th to the 16th) which tell about the birtplace. It is also specified whether the customer is recognised as student, but this information has been ignored because some customers classified as students have a too high age.
  - *codici_validazione* describes the meaning of each type of validation: for passengers with transport credit, it is specified if they are boarding or

alighting; for other passengers, there is just an indication that the travel document has been correctly validated.

- **Public transport offer**. This category includes the list of the providers of the service, its timetable, the structure of the network, the fares and the features of each stop. Specifically, data are organised according to the GTFS standard, which includes the following mandatory tables and attributes:

  - *agency* contains code (optional), name, url and timezone of each transport agency.
  - *routes* tells about code, name and means of transportation of each route, but it includes also some additional data, such as the agency which executes it.
  - *trips* (a trip is an instance of a route, usually unique during a day) reports the code of the trip, of the associate route and of the service, which allows to link to the *calendar_dates* table (see below). There are also other attributes: the next terminal stop, the name, the direction (useful for bidirectional trips on the same route) and the shape it belongs to (see among the optional tables).
  - *calendar_dates* has a list of service codes and dates: by comparing the code with that present in the *trips* table, the dates in which a trip is executed can be retrieved. It replaces the *calendar* table, which is usually mandatory.
  - *stops* includes code, name and geographical coordinates of each stop of the network, but in this case also the zone in which the stop is located and whether it's a station (a group of stops, rather than a single one) or not are available.
  - *stop_times* tells, for each couple (*trip*, *stop*) the planned arrival and departure times of the trip at that stop and the ranking of the stop in the trip. In this case, there are also the next terminal from that stop and two codes which mean whether the passengers should contact the agency or the driver when they want to board (*pickup_type*) or alight (*drop_off_type*) at that stop.

  There is also an optional table, named *shapes*, which has a code linked to the trips and for each of these, there is a list of geographical coordinates and ranking of the points that form the shape. The relationships among the seven tables are represented in Figure 3.3: in each table, the primary key is marked with PK, while each arrow goes from the external key to the primary, representing a relationship of kind many-to-one.

- **Weather conditions**. They must be taken into account, since, when the weather is bad, some categories of people may be more willing to use public transport (such as students, since many of them don't have an alternative), while other ones, for example people who can delay their travel, may be discouraged. [**34**] provides the daily historical weather for each Italian municipality, so it can be a good source.

- **List of municipalities**. For the purpose of a faster searching of the weather data, it's useful to have a list of all the possible places, so to link a stop point with its historical weather: this is downloaded from [**35**]. The research is of course limited to the six provinces of interest.

- **Geographical coordinates of the municipalities**. When the stop name is not meaningful, it's necessary to find the nearest municipality, so the OpenCage Geocoding API, available at [**36**], is used to attach the latitude and longitude to each of the places found at the previous point.

- **Geographical and demographical data of the census sections**. They can be downloaded from [**38**]: the most recent information refer to the last census, which dates back to 2011. For each census section, the relevant information (not all used here) are the area of the surface, the population, the number of students, employed people, buildings for living and for other purposes, but there are dozens of other values, concerning also houses, families and foreigners.

- **Lists of codes of Italian municipalities and foreign countries**. They are respectively taken from [**39**] and [**40**] and allow to link the fiscal code of a customer to his birthplace.

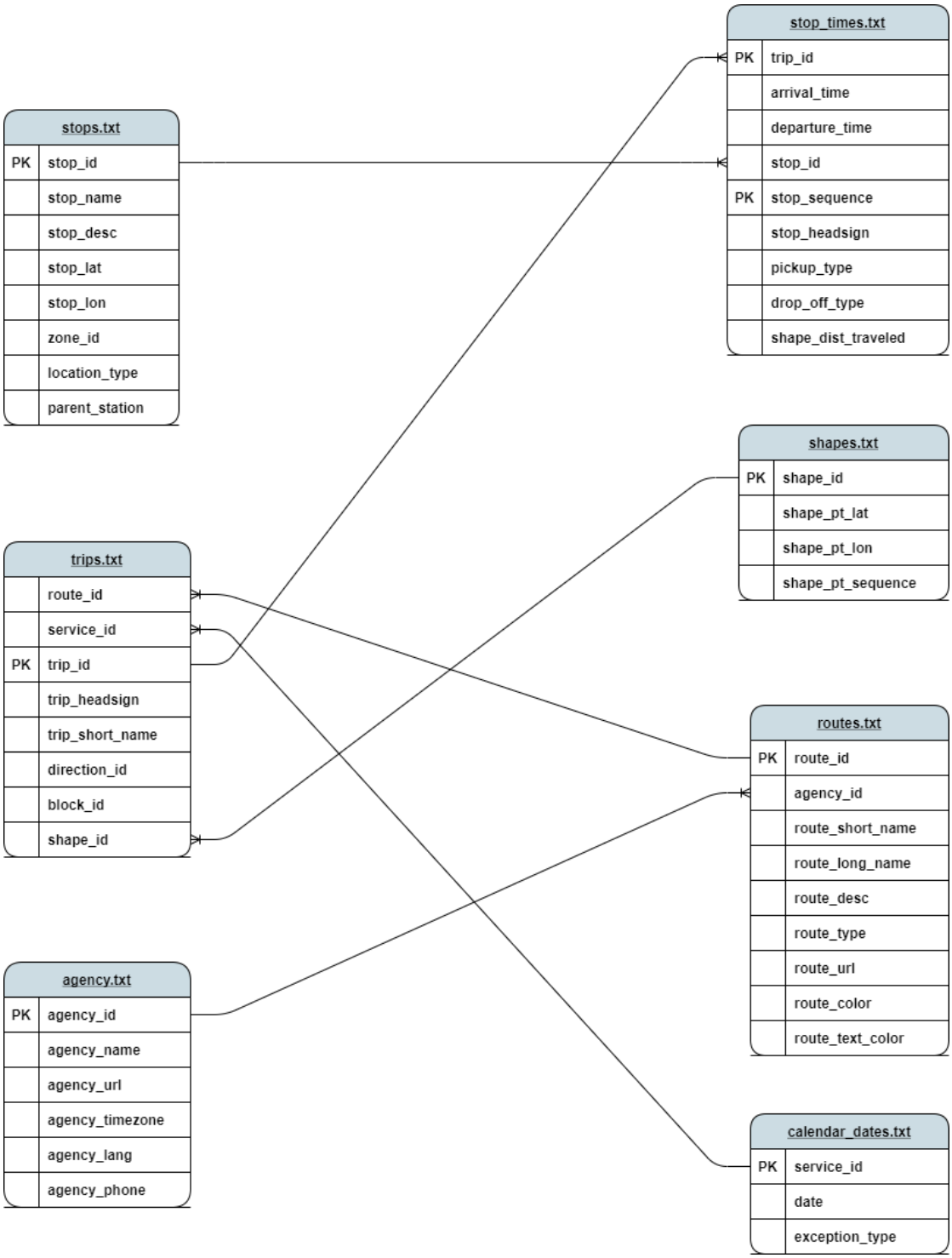

Figure 3.3: Entity-relationship diagram for GTFS data about public transport offer.

Chapter 4

Preprocessing

4.1 Data cleaning

Cleaning data consists in removing all the records that clearly look wrong. These errors may have been generated while extracting data from their source, or they may result from inconsistencies between different sources, or, even before, there have been some malfunctions of the device which has recorded them. Anyway, detecting and properly managing any type of anomaly is fundamental for all the following steps of the analysis, because also the smallest error may be then amplified and finally produce wrong results.

The first concern is with duplicated validations. These may occur with tickets bought on board: since the timestamp is recorded with precision up to the minutes and different passengers can't be distinguished in this case, there may be two records with the same timestamp, stop point, vehicle, route and trip. On the other hand, they are not allowed in the validations table, where also the user code is available. However, no pair of duplicate validations has been found in this table.

Another important issue is given by the records without the stop point: this is probably the most important information; also when the route and/or the trip are available, it is difficult to estimate it. Therefore, rows where the stop is missing have been removed, even if, in theory, some statistics which don't have to do with the stop points could have been computed anyway.

The most evident problem encountered at this point is that there are many pairs user+day with an incredible number of validations: they can't be treated with certainty as errors, but it is worth to deepen this behaviour. Since it seems that some of these many validations occur in few minutes but very far one from another, it has been useful to estimate the speed that the user (and, as a consequence, the bus on which he travels) would need to make both validations. This is because, even if the travel document is impersonal (so it may pass from a person to another),

it is however *unique*, so it can't be found in very far places at the same time, or with a difference of few minutes.

In order to find pairs of validations with this problem, the mean speed of each trip has at first been estimated: the tables *stop_times* and *stops* have been joined, so to have the sequence of the geographical coordinates of the stops touched by each trip. The distance between stops has been computed in an approximated way, using the formula for the Euclidean distance on a plane: since the area of interest is relatively small, the curvature of the Earth doesn't affect very much the accuracy of the result. In order to change the unit measures from degrees to kilometres, which are definitely more familiar, the result has been multiplied by 111, which approximates the distance between two meridians (111.32 km) and two parallels (110.95 km), if the Earth is supposed to be divided into 360 meridians and 180 parallels. Of course, this has required to assume that the bus draws straight lines between the stops, so the distance has been under-estimated, and therefore also the speed. The duration of the trip between any two consecutive stops has instead been computed as the differences between the respective arrival or departure times (they are always equal): in order to express the speed in kilometer per hour, duration have been converted to a decimal base (for example, one hour and thirty minutes corresponds to a duration of 1.5). However, since in the table *stop_times* the times are rounded to minutes (also seconds are shown, but they are always null), in some cases they are equal for two consecutive stops: this would have given null duration and thus infinite speed, which would have made not finite also the average. Therefore, journeys between these kind of stops have been excluded from the computation: they have been associated to a missing value, which, when computing the mean of all the speeds for a certain trip, has been ignored. The obtained table has been merged with the main dataset (based on the trip column): since there are some validations without the trip indication, also the mean speed trip is missing in the corresponding row. Therefore, the mean speed has been set to a prudent threshold, such as 80 km/h.

The same computation has been performed for the theoretical speed of each customer: first of all, validations have been sorted mainly by user and secondarily by timestamp (containing both day and hour). In this case, two columns have been created: the speed *before* (which relates the current validation to the previous one in the ordering) and the speed *after* (referred to the interval between the current and the following validation). The computation has been possible only with at least two validations of the same person in the same day: if this is not the case, the speed has been set to 0, so to be sure that the record would have been preserved. Also when there is an alighting validation, the speed *after* of that row and the speed *before* of the following row have been set to 0, since the user may have moved with any other means of transportation in the meantime, so it would have been not careful to compare his speed with that estimated for a public

bus service. It should be clear that, at the end, the two columns of the speed are very similar (the second can be computed from the first by shifting it up by one position and the starting value of the first corresponds to the final value of the second). However, using two columns for the speed is necessary because validations must be treated differently, depending on their type: for a boarding, the trip speed has been compared with the speed of the user *after*; for an alighting, it has been compared with the speed *before*. Moreover, it should be remarked that, in this case, the mean speed is even more under-estimated than in the case of trips: first of all, because a customer usually travels for more than one stop, so the approximation of the true path with a straight line is even more rough; second, the duration of the trip is usually over-estimated (unless the first is a boarding and the second is an alighting), since the interval between two validations includes also time spent on waiting for the second bus. For these reasons, removing all the rows where the user speed (*before* or *after*, depending on the cases) is greater than the mean speed of the trip has been precautionary: it is likely to preserve all the rows without errors, but it does not grant to remove all the errors. For what concerns the rows without the trip code, only the user speed has been estimated and it has been compared with the threshold discussed above (80 km/h). The distribution of the mean speed of the customers computed in this way (specifically, the speed *after*), after having removed the supposed wrong records, is shown in Figure 4.1, where, for a better visualization, the vertical axes is logarithmic.

However, removing validations too close in time has only partially solved the problem of too many records in a single day related to the same customer: before this operation, some users had been associated to about 30 validations in a single day. Then, the maximal value found was 14: however, it was still too much. These codes may represent test users or ticket inspectors. Personal and impersonal user codes have been treated differently: for the first kind (which are more likely to correspond to true customers) only the validations related to a pair (user+day) with more than ten records have been removed; in the latter case, all the records related to that user (also when they are less than ten in the same day) have been removed. However, also eight or nine validations in a day are unlikely, but is difficult to distinguish validations of true customers from those made by other people.

For what concerns users, another oddity, though involving a smaller number of records, has been represented by those classified as retired people, even if they were relatively young: specifically, this has happened for 5 users who were less than 42 years old. They were associated to few dozens of validations, which have been therefore removed.

In view of the segmentation of the stop points performed in Chapter 6, and for the computation of certain statistics in the following one, also some other records have been removed:

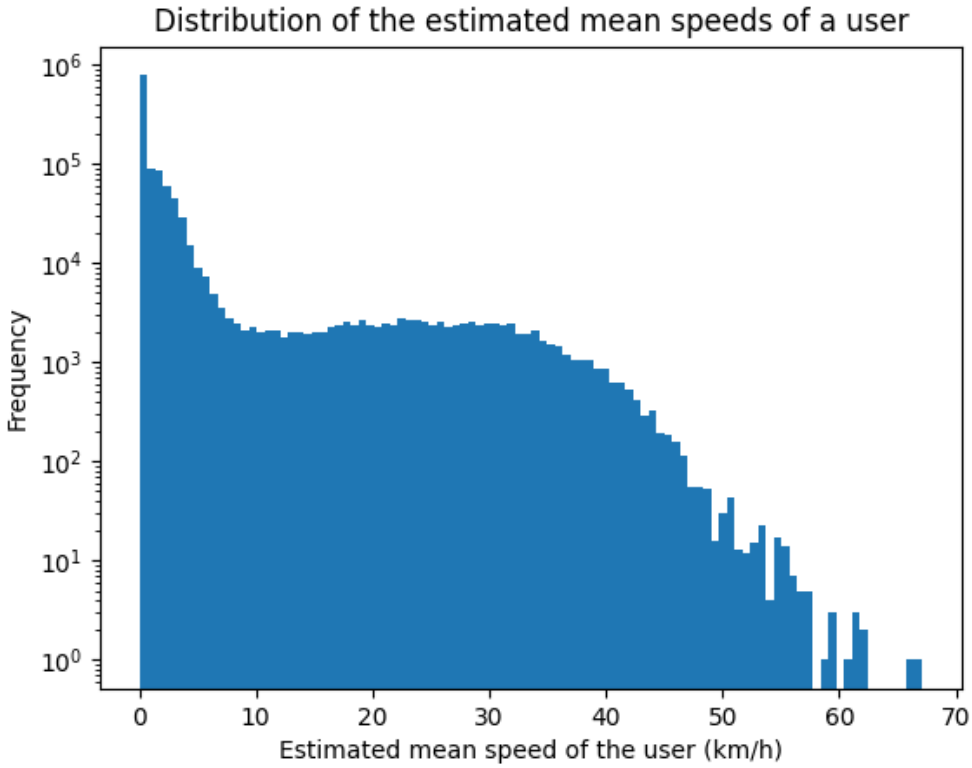

Figure 4.1: Distribution of the estimated mean speed of the users between two consecutive validations occurring on the same day, after having removed the supposed errors.

- Since the analysis has been often focused on boarding validations, those related to an alighting (available only for passengers with transport credit) have been discarded
- When customers buy the ticket on board, since the fare depends on the towns of departure and arrival (and not on the exact stop points), the stop point is not recorded with precision, but a default stop point is used for each town. This would be an error when computing statistics about the single stop points or when clustering the stops themselves. To avoid it, it has been sufficient to remove rows without user code, because they are all and the only records with this characteristic.
- Records without the trip code have been usually preserved, except when computing statistics related to trips themselves. Actually, one could try to estimate the trip by considering the validations occurring at the same stop

point and at the same time (with a tolerance of few minutes). However, in most of the cases all the validations related to the same stop point and close in time have the trip missing. This could have been expected, since the missing of the route is likely to depend on the vehicle, rather than the user or his smart card.

- Records without the route code may have affected the segmentation of the stop points, so they have been removed at that moment, due to a reasoning similar to what explained for the trips.

The followings are other anomalies which have been ignored, in the sense that the related validations have not been removed, for the reasons that will be soon explained:

- Few records contain an association between stop point and trip which doesn't appear in the table *stop_times* of the GTFS, as if the trip is not planned to meet that stop. As expected, this usually happens when the trip or the user code is missing (the second is the case when the ticket is bought on board, with the consequences just explained); however, there are other few dozens of records with this problem. For them, it is likely that the error comes from the GTFS tables, since these are usually generated with the purpose of planning the service in the future. On the contrary, in this case the tables have been extracted after the reference period: if in the meantime the service has been modified or just some coding has changed, it is likely to find some inconsistencies.

- Some trips appear in the *stop_times* table, but not in *trips*. Of course, this problem has nothing to do with the validations.

- Based on the tables *calendar_dates*, *trips* and *stop_times* of the GTFS, some stop points are associated to validations also in days when there isn't any trip planned to pass by them. However, also this phenomenon is very rare and has been attributed to errors in the GTFS.

- There is a shape (in the homonym table) formed by some points whose geographical coordinates are clearly wrong, since the shape immediately jumps from Piemonte to Sicilia (more than 1000 kilometers apart) and then goes back across the whole Italy. However, the trips referring to this shape look correct, since they link two places in the expected territory.

- Some validations are related to users with more than 100 years (nearly up to 120). This may be due to the fact that the true customer is different from the owner of the travel document, also because these very old users are not recognised as retired people. Such a classification, as explained in the next

Chapter, is based on the table *titoli*, which is more reliable than the table *utenti*. In other words, maybe it doesn't matter if someone uses the smart card of another person, but he shouldn't be allowed to charge a travel document for which he doesn't have the right.

- Finally, for the customers with transport credit, there is a great difference between the number of boarding and alighting validations. This happens because, in some towns, also these travellers don't have to validate when they arrive at destination.

4.2 Quality of data

Even by removing also validations without route, trip or user code (which, as already said, for some analysis are retained), the quality of the dataset looks quite good. Actually, in Figure 4.2 the top bar has been divided into four sectors, corresponding to how the errors, anomalies, or just annoying records discussed in the previous section have been tackled. Each pie chart shows the proportion of each type of anomaly in each group. Some of the errors discussed above are not present in the pie chart, either because they are difficult to quantify or because they don't involve the validations table. This shows that more than 3/4 of the records look fully correct and, in the majority of the cases, about 90% of them (green, yellow and possibly part of the orange sector) has been useful.

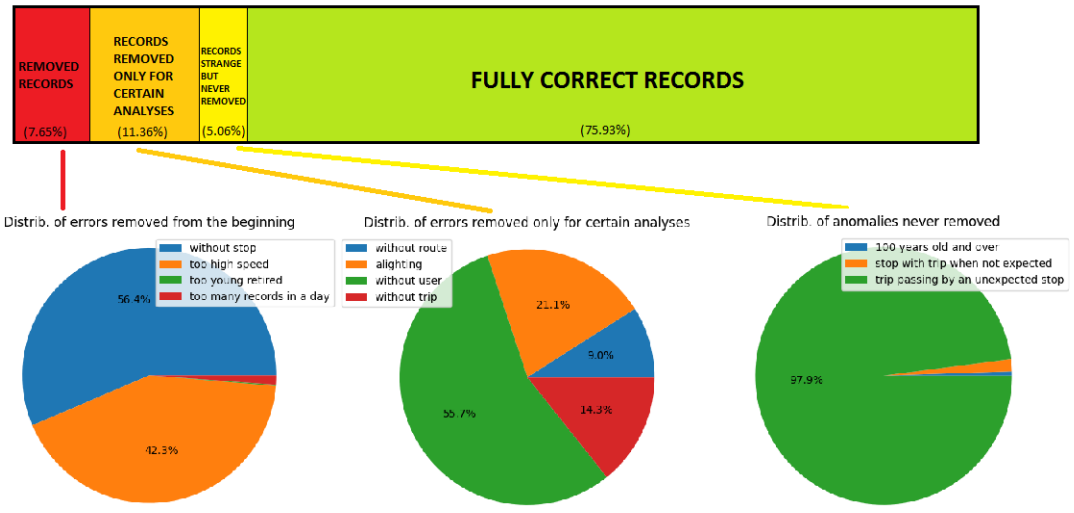

Figure 4.2: Distribution of the errors in the validations table according to their category.

It should be highlighted that the computed frequencies of each kind of error

may vary, depending on the order in which the operations are performed, which is the same as they have been listed in the previous section. Therefore, percentages in the top bar of the figure should be intended as the frequencies of the corresponding groups with respect to the original number of records, but in the reality these values may be slightly higher, since some of the records with that problem may have already been removed. Another remark is related to the percentages shown in the pie chart: in this case, the sample is formed by all the records belonging to that group, but the slices of the pie may partially overlap, if some records have more than one of the issues reported in the legend. However, the reported values tell the frequency of the corresponding problem in the whole group. Moreover, as already said, not all the problems discussed in the previous section have been reported: however, they would have been classified among the "ignored" anomalies (the yellow group), so they have not been removed. Specifically, some of them are related only to the GTFS tables (trip appearing in *stop_times* but not in *trips*, shape with strange coordinates). The last one is the great unbalancing between boarding and alighting validations of customers with transport credit, due, as already said, to the alighting not being mandatory in some towns. Unfortunately, it is difficult to quantify the wrong records (if present), since it can't be easily identified the reason of the missing alighting validation. The last remark is related to the records with an unexpected association between stop point and trip, which represents the great majority of the yellow group: actually, in many of these records either the trip or the user code is missing (the latter case corresponds to tickets bought on board, so to a possibly inexact stop point), thus the association would however have been wrong or not possible. If all these records were discarded, only few dozens of validations would have been left with this problem.

Chapter 5

Elaboration and analysis of the input data

5.1 Joining data in a single table

In Python, there's a package, named *gtfs_kit*, which includes some functions that automatically analyze the tables about the public transport offer. A full explanation can be found in the documentation of the package (available at [**37**]); those used in this case have been:

- *get_trips* gives the *trips* table
- *compute_trip_stats* computes some statistics for each trip, optionally restricting to the specified routes
- *get_routes* gives the *routes* table
- *compute_route_stats* returns some statistics for the routes, by considering only the specified trips and dates: in this case, all of them are included.
- *get_stops* gives the *stops* table
- *compute_stop_stats* computes some statistics for the stops, related to each of the specified dates (the whole month of October) and optionally to a subset of stops and to a time interval. In this case, the number of routes and trips visiting the stop during each day will be useful.
- *geometrize_stops* turns the *stops* table into a GeoDataFrame, which means that each stop can be visualized as a point on a geographical map, by using a suitable software (QGis in this case). The map can be made clearer: for example, it can be seen where the stop points are located and the color

of the markers changes based on the total number of incoming validations. Moreover, a *spatial join* is performed to add (see later) information about the census section to which the stop point belongs. Since each point on the map corresponds to a stop and at the same time the map can be divided into polygons representing the census sections (not represented in 3.2), this operation consists in associating each point to the polygon which contains it.

- *get_stop_times*, which returns the *stop_times* table.

Some useful information have been extracted also from the tables about the public transport demand. Rows without the stop point, which will be quite useless in the following analysis, have been, as already said, excluded. The tables *validazioni* and *biglietti* have been concatenated, since they have the same scheme (also the latter contains the column *codice_utente*, but it's completely empty). The resulting table has been reorganised as follows: the temporal attribute is split, so to have day and time in two separate columns. The time itself has been split, so to highlight hours and minutes (useful in the following analysis). Further information about the stop where the validation occurs have been collected from the table *stops*, which has been merged with the main table.

Moreover, for each row, some new columns have been created:

- **weekday**. For working days, this has been computed as expected, but starting from 0 (0 is associated with Monday, 1 with Tuesday, 2 with Wednesday, 3 with Thursday, 4 with Friday). Saturdays have been coded with 5 and Sundays with 6. Then, a holiday which is not on Sunday has been however coded with 6, while half-holidays like school holidays or typical days of summer break have been associated to 5. To create an association between half-holidays and number 5 and between holidays and number 6, two lists containing these special days have been created and filled: since these days are specific for each country, they must be provided by the user. However, October 2019 doesn't include half-holidays or holidays different from Saturdays or Sundays respectively, but this methodology may be generally useful when considering a larger time interval.

- **data_stringa**. It's just the date, written in a string format (YYYYMMDD).

- **giorno_mese**. Starting from the previous point, if the day of the month ranges from 1 to 9, the first digit (0) has been removed: this operation has been necessary for a correct merging with the historical weather table (see later).

- **luogo**. For each stop, it corresponds to the first part of its name: it gives an idea of the location, but it has been useful also when collecting historical weather data, for a faster association between a stop and its location.

It has been useful to concatenate, even if something is missing, the stop and the trip with the corresponding route. Moreover, a secondary table without rows where the trip lacks has been created.

Also the weather conditions recorded in the location has been added to the validations table. Unfortunately, historical data for each single timestamp and stop are not available in [**34**]: however, a good level of detail has been reached, since they have been collected for each municipality located in one of the six provinces of interest and for each day of 2019. For each of them, the geographical coordinates have been attached. Then, the location of each stop has been read: if it corresponded to one of the recorded towns, the historical weather for each day of the period of interest have been straightly recorded. If instead the first part of the stop name was not meaningful (since it for example referred to a street, without the town name), the historical weather has been searched for by using the nearest town among those in the list. The distance between the stop and each town is computed in an approximated way, using again the formula for the Euclidean distance on a plane: moreover, in this case the interest was just on the ranking of the distances, while the absolute values didn't matter, so the square root and the scale factor which turns coordinates differences into kilometric ones (111) have been discarded, because they didn't change the ranking. Therefore, the unitary distance was that between two points with the longitudes differing by one degree and same latitude (or viceversa). In both cases, the association between each stop and the location used for looking for the historical weather has been saved in a dictionary, a common data structure consisting in several pairs whose members are respectively called *key* and *value*. In this way, if the same stop appeared more than once in the validations table, the research of the nearest town (if necessary) and of the weather has been performed only the first time. Since this is usually the case (it's likely that a stop appears more than once during a month), historical weather data for that town have been collected for the whole year, not depending on the day when the validation occurs. For each couple (*day*, *town*), where the day includes also the month (precisely, it's the attribute *giorno_mese* previously created), the following information have been made available:

- minimum temperature
- maximum temperature
- amount of rain (measured in mm)
- general weather condition (sunny, a little cloudy, cloudy, changeable, rainy, snowy, foggy, stormy, rain mixed with snow)

Of course, the weather during a day may considerably change: the label *changeable* is useful to represent this situation. A missing information which could be relevant

(as explained in [**22**]) is the wind speed: however, it is not very useful at daily level, since it can change very fast. All the data have been put in a single table: each row is identified by the pair (*giorno_mese*, *location*). The location is that of the stop, therefore this table has been joined, based on these two attributes, with that containing the validations.

The result has been in turn merged with other tables, so to have a full overview. Specifically, it shared the attribute *titolo_viaggio* with the table *titoli* and the attribute *codice_utente* with the table *utenti*.

Also, some other information have been added to the validations table:

- Based on the category of customer, the desired splitting is into students, retired people and other ones. The category has been read either from the tables *utenti* or *titoli*, depending con the cases. Actually, the latter looked to provide all the needed information (students and retired were explicitly classified, the third category could be found by subtraction), but in this way a lot of relatively old people were classified as students, which was obviously a mistake. Instead, in the table *utenti* all the customers recognised as students were at most 25 years old, so for this category it was better trusting on the table *utenti*. Moreover, this is not a subset of the students as found in the table. Finally, all the users with a code (even if impersonal) and not put in the previous categories have been classified as *others*.

- The result has been merged with the table generated by the function *compute_stop_stats*, based on the attributes *stop_id* and *date*. Also in this case, the association was sometimes missing, due to the fact that this function is based on the GTFS tables *calendar_dates*, *stop_times* and *trips*, and, as already said, some validations occurred at stop points where, on that day, they were unexpected.

- Given the table *stop_times*, a binary variable which tells if that stop is (1) or not (0) the first for that trip is attached to each row; then, this information has been added also to the validations table, in the new column *terminal*. Unfortunately, as explained in Section 4.1, also this attribute wasn't be always available, for example when the trip code was unknown or for rows coming from the table *biglietti*, but also when the association between stop point and trip didn't appear in the table *stop_times*.

- After having linked each stop point with the census section where it is located, the result has been in turn joined with another table, containing all the useful data about each census section of the area of interest. This table has been obtained by merging three other tables one with another: these respectively involved the geographical, demographical and economical features of the six provinces of interest. One of these indicator, the population density of each

census section, was not explicitly available but it could be easily computed as the ratio between total population and area of the section: as a consequence, it has been expressed as number of inhabitants per square meter. An important attribute, which will be used also when segmenting the stop points, has been the kind of place (*TIPO_LOC*), which can assume 4 different values: 1 stands for an automonous group of buildings, usually including houses and facilities; 2 denotes a smaller group of houses, without other types of buildings; 3 indicates a place where the production activities prevail; finally, 4 stands for some scattered houses. However, this is a property of the locality (an intermediate level between the municipality and the census section): this means that each section belonging to a certain locality has the same value for this attribute.

- Finally, the available characters of the fiscal code of the user have been translated into a more meaningful information. The column *NAZ_CODFISC* contain the last 5 digits of the fiscal code: apart from the last one, they are related to the place of birth of the user. The first letter told if he was foreigner (if it is Z) or not; together with the three following digits, it allowed to identify the municipality (or, in case of foreigners, the country) where he was born. The association between the code and the birthplace is read from [**F**] in case of Italian municipalities and from [**G**] for foreign countries. Thus, the birthplace has replaced the column *NAZ_CODFISC*.

5.2 Estimation of the destination

Except for passengers validating with transport credit, only the boarding stop of a user was known, while the alighting one had to be estimated from the available data: this has been done through the so called *virtual check-out* algorithm. For this purpose, the tables about demand and offer have been reorganized in a more efficient way. Specifically, nine new data structures have been built. In general, they are nested dictionaries:

- **corse** is entirely based on the table *trips*: each trip code is a key and the associated value is formed by the concatenation of the code itself, the shape to which it belongs, the route and the direction.

- **calendario** has again a trip as key, but the value is the list of dates in which that trip is planned to be executed. This has been further reorganized so to have also the number of these dates.

- **codici_fermate** trivially associates each stop with itself; in other contexts, it would have been useful to link the coding used for stop points with that used for pairs stop point+agency

- **corse_esercizi** is a nested dictionary: the outer has the trip as key, while the inner has each visited stop as a key and some other information as value: the date of execution (read from *calendario*), trip, route, shape, direction (all available in *corsa*), the ranking of the stop in the trip, the stop code and the planned time of arrival (based on the table *stop_times*).

- **fermate** contains all the stop points: the values consist in stop code, name, latitude and longitude, everything taken from the table *stops*.

- **tratte** tells about all the links between two stops: the code (in this case, just the couple of stops), origin, destination and covered distance.

- **corse_consuntivo** just reports a list of the executed trips associated with the planned one: similarly to *codici_fermate*, in this case key and value are always the same

- **utenti_check** has as keys all the users which have validated at least once with transport credit: the values are nested dictionaries with the dates of these journeys as keys and, for each date, the list of validations (timestamp, user code, trip, stop point and outcome, which may be ckeck-in or check-out). These information have been taken from the table *validazioni*: however, the stop point and the trip were sometimes missing.

- **utenti** is very similar to the previous one: the only differences are the customers involved (those travelling at least once not with transport credit), the outcome (only the check-in is recorded) and the source table, since in this case also the validations in the table *biglietti* must be taken into account.

Figure 5.1 shows the entity-relationship diagram among these tables: the primary key of each is marked with PK, while the arrows highlight the relationships, with equal cardinality (many-to-many or one-to-one) if bidirectional, many-to-one if unidirectional.

The next step, in view of the prediction of the demand, has been the estimation of the destination for passengers which don't have to validate when they get off from the bus: since this operation is strongly based on the boarding stop, the records without this information have been again discarded. Moreover, other checks similar to those performed to find anomalies in the previous Chapter, have been done. Unfortunately, in more than half of the cases, this estimation has failed: this can be due to errors in the stop coding, or to a great distance between the supposed and the true stop. Moreover, since the estimation is strongly based on the origin of the following journey made by the same user on the same day, the task has been very difficult when there was only one trip per day.

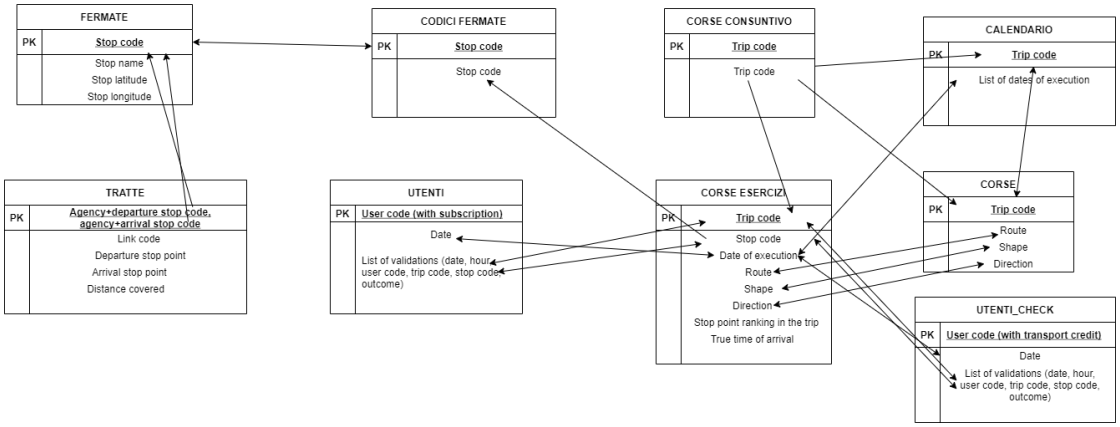

Figure 5.1: Entity-relationship diagram for the nine tables used to estimate the destination.

5.3 Analysis of the data

By considering only customers with at least a validation during October 2019 and not with an impersonal document, most of them were young people, which will be clear from the distribution of travel documents and users features: subscriptions for students represent almost half of the travel documents, while just one every six is a single or a group of tickets. Indeed, almost 3/4 of the users were identified as students and, by restricting to users whose age and gender were known, more than two thirds were less than 20 years old. In this age range, men were more than women, while in the other cases (and in general) women prevailed, with the difference getting bigger as users became elder. Young people had also the tendency to travel more than other ones, since their frequency (and that of the related travel documents) was even bigger by considering the validations table, where each user's weight corresponded to the number of his validations. In general, less than 10% of the validations were done with transport credit: this indeed means that a smaller fraction of users travel in this way, since, for each of their journeys, both the boarding and the alighting were recorded, at least in theory. The analysis of the public transport demand has been performed, depending on the cases, by considering one or both of the following indicators:

- Percentage of validations, with respect to a certain sample.
- Mean daily number of validations, especially when it is necessary to "normalize" the values in order to compare them.

The analysis has been focused on the following dimensions:

- **Hour**. The day has been divided into 24 timeslots (one hour each), with the first starting at midnight. In the charts, timeslot n always stands for the interval between hours n and $n+1$ (for example, timeslot 7 represents times from 7:00 to 8:00)
- **Weekday**. This is intended as explained above, but in the charts the numbers have been replaced by the first three letters of the weekday's name.
- **Type of day**. This has been just a regrouping of the previous type of partition: working days have been joined together in the first group, while half-holidays and holidays have formed the second and third group, respectively.
- **Day of month**. For a more detailed analysis, all the days of the sample month have been compared.
- **Stop point**. Each stop point with at least an associated validation has been considered, usually with the associated route, too (since some stops may be shared by more than one route). Due to what said above, rows coming from the table *biglietti* (those without user code) have been ignored. Since there are thousands of stops, in some cases only the most crowded have been included in the analysis, so to make it more clear.
- **Trip**. Similarly to stops, each trip with at least a validation has been taken into account, together with the route it belongs to: in this case, even if each trip always relates to a single route, this has been useful to better understand the most crowded routes, since different trips associated to the same route in a single day have different codes.
- **Location**. It's based on the stop name, whose first part often tells about where the stop is located: in this way, neighbour stops have been in some sense joined together. However, not all the stop names are meaningful for this purpose.
- **Travel document**. Passengers have been partitioned according to the kind of travel document used in the validation, which could be read from the columns *TIPOLOGIA* and *TIPOLOGIAPERIODICITA* of the table *titoli*: they have been both used to split into ticket, carnet of tickets, subscription for students, other kind of subscription, while, in order to find the length of the period of validity of the document (this is not always well defined, such as for a single ticket), which ranges from one week to one year, the latter is sufficient.
- **Category of passenger**. The column *categoria* in the table *utenti* has allowed to split into students, retired and other people. Sometimes, there has been a further split according to the outcome of the validation, which told

whether the customer had boarded with transport credit or not (alighting validations have been already removed).

- **Demographical data**. People have been split also according to gender and age range (until 19 years, from 20 to 30, from 31 to 41, from 42 to 52, from 53 to 64, from 65 on). This has been very helpful for *other* users, but not for students and retired people, since their age has low variance and their gender has low significance.
- **Weather situation**. Data have been grouped according to the weather condition recorded during that day in that specific location.

The statistics listed below are computed: they are equipped with the charts shown in Chapter 7. A validation is intended as one or more passengers boarding, so the validations table has been used.

- Some pie graphs have given a first glance of the features of the passengers: these have been based on the validations table (therefore, each user appears as many times as the number of validations associated to him) in Figure 7.2, while in Figure 7.3 these have been built according to the tables *utenti* and *titoli*. In the first figure, the two top graphs are based only on validation with a known user code. In the second one, all the users and travel documents associated to at least one validation appear exactly once, unless the plotted features are unknown.
- Percentage of validations (with respect to the total), mean daily number by weekday and percentage by day of the sample month, always split by timeslot (Figure 7.1).
- Percentage of validations at each stop point and at each location, extracted by the stop name as explained above: the most frequent stops and locations have been labelled and the stops have been also associated with the route (Figure 7.10).
- Percentage of validations for each trip: the 14 most recurrent trips have been labelled and associated with the route. Moreover, these 14 trips have been divided into two groups and, for each group, the mean number of validations by weekday has been shown separately (Figure 7.11).
- The stops with the highest mean daily number of validations were likely to have also a big variance and the same holds for the trips and the users: therefore, for the first 14 stops (respectively, trips and users) in the ranking, individual boxplots have been drawn in Figure 7.12. For stop points and users, only records with the user code have been considered; for the trips, only records with the trip code have been considered.

- Percentage of validations for each category of user (with respect to the total for that type of day), also bisected on the type of validation (with transport credit or not), always split by timeslot and type of day (Figure 7.6).
- Mean daily number of validations by weekday and category of user, always split by timeslot (Figure 7.7).
- In order to better know the habits of *other* users (one of the categories considered in the two previous statistics), these have been further divided by age: (intervals with low and high values of age are wider because many of their members were supposed of having been already classified as students or retired, respectively) and gender (men on the left column of Figure 7.8, women on the right column). The percentage of validations for each group has been shown, divided by type of day and timeslot.
- Percentage of validations, with respect to the total for that type of day, by kind of travel document, separately for each type of day and always split by timeslot (Figure 7.9).
- Mean daily number of validations by weather condition and type of day, split by timeslot (Figure 7.13).

Chapter 6

Characterization of the stop points

As anticipated in the previous sections, records without route code, user code (where the stop point may be inexact) or related to alightings (since the focus is on incoming demand) have been fully discarded for this analysis, together with all the validations already removed at the beginning of Chapter 4 (the red group in the top bar of Figure 4.2). All of them were records which may have affected the efficiency of the clustering algorithms used for trying to characterize the single stop points. Moreover, since the segmentation has been based only on temporal and possibly geographical variables, some of the attributes computed in the previous section for the analysis of the demand (such as features of customers and travel documents, weather conditions) have been discarded: in practice, at the beginning, the retained variables have been stop+route, timestamp, day, longitude and latitude of the stop, weekday and number of people. Then, as explained in Chapter 7, also some other variables have been restored when describing each group.

6.1 Discretization of time

A preliminary operation before the true clustering has consisted in choosing the most suitable size of the bins used to discretize time. Actually, for each validation, the related timestamp has been replaced by the time bin it belongs to, in order to associate records which are close in time (and, obviously, with the same stop point and possibly route). However, deciding the size of the bins has been not easy: of course, as it increased, the number of bins, and therefore also time and memory consumption, was reduced; at the same time, records related to very far timestamps may have been put together: this would ultimately have affected the efficiency of the clustering. Before taking a decision, some trials have been done: in this case,

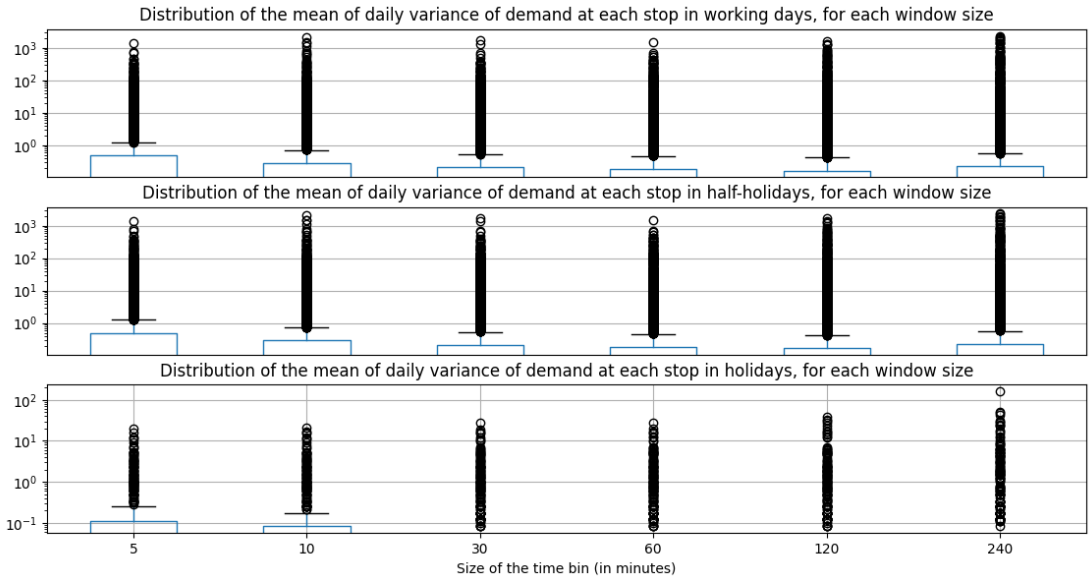

Figure 6.1: Distribution of the mean daily variance of the demand, at each stop point, for each type of day and size of the time bins.

six possible reasonable sizes have been tried (5, 10, 30, 60, 120 and 240 minutes). Moreover, what follows has been separately done for working days (from Monday to Friday), half-holidays (like Saturday) and holidays, as having seen (details have been given in the following Chapter) that the trend of the demand sensibly changed among them. Therefore, the attribute *type of day* has replaced the weekday from now on. Then, for each pair stop+route, the demand at each time bin, during the whole period (October 2019) has been computed: of course, the demand is the sum of the attribute *NUM_PERSONE* over all the records falling in that time bin, so it corresponded to the number of people who had boarded on the bus during that time interval. Then, a sort of variance of the demand over the different time bins of a single day has been computed, separately for each pair stop+route and day. Actually, the computation of the variance has been restricted to the time bins where at least one validation occurred at the considered stop+route. This has been done because, for many stops, there were very few time bins with a demand different from 0, for example during the night; thus, considering all the zeros in the computation would have probably increased the variance more than desired. In other words, the variance of two stop points, one with about 5 validations per time bin during the day, and the other with about 20, should have been similar. Finally, the mean of these variances has been computed, so to get, for each pair, a mean variance of the demand for each of the three types of day. This has been repeated for each possible size of the time bins, so to get, for each size and type of day, n

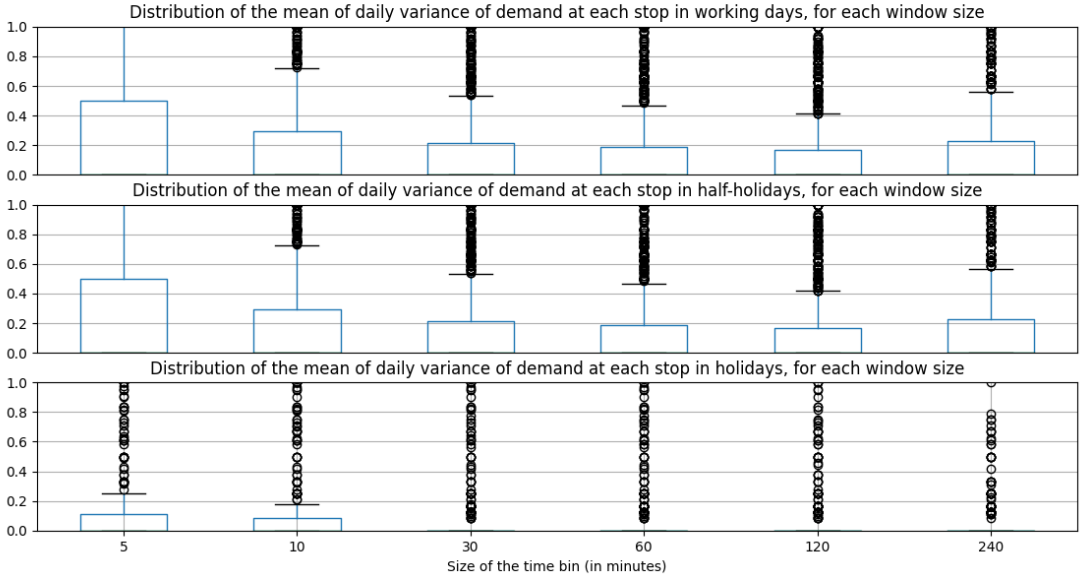

Figure 6.2: Zoom of the distribution of the mean daily variance of the demand, at each stop point, for each type of day and size of the time bins.

values of variance, where n is the number of different pairs stop+route. It was likely that the distribution of these values would approach 0 as the size increased: however, the strategy of finding a good compromise between clustering performance and computational complexity has been pursued by looking for an elbow in the relationship between the distribution of the mean variances and the sizes of the time bins. For a better visualization, the first variable has been shown in 18 boxplots (one for each time bin and type of day), divided into three charts (one for each type of day), reported in Figure 6.1. Each boxplot shows the distribution of the mean daily variance of the demand, for a certain type of day (written in the title) and size of time bin (specified on the horizontal axis), across the pairs stop+route. Each box extends from the lower to the upper quartile values of the data, with a green line at the median; the whiskers extend from the box to show the range of the data. Flier points are those past the end of the whiskers. For a better visualization, the vertical axes has a logarithmic scale. However, due to some very far outliers, the boxes are very pushed towards the bottom of the charts, so they can be seen more clearly in the Figure 6.2, where the boxes are zoomed on values between 0 and 1 and their boundaries are more visible. Of course, this representation leaves out all the stop points with mean variance greater than one, but it makes more evident that the distribution of the mean variances is not monotonically decreasing with the

size of the time bin: 120 and 240 minutes should certainly be discarded. The doubt was between 30 and 60 minutes: however, the latter is chosen. The improvement between the two sizes is small, but the choice has also a computational reason. In the previous Chapter, the hours and the minutes had been extracted from each timestamp and this operation has proved to be rather faster than discretizing time by using some user-specified intervals. Therefore, choosing 60 minutes as size has allowed just to replace each timestamp with the hour number already computed.

6.2 Clustering techniques

In order to choose the best clustering technique, even before tuning the parameters of a certain algorithm, such an algorithm must be carefully selected. Actually, each of them as its own advantages and disadvantages, so the best compromise should be found. First of all, only techniques which performs well with a quite big number of samples have been selected for comparison. In this scenario, there are three main groups of algorithms:

- **k-means** needs the desired number of groups to be specified. Such a number of points are chosen as starting members of the groups; all the other samples are assigned to the nearest of these points. From now on, at each iteration the centroid of each group is computed and all the points are assigned to the group with the nearest centroid. In this way, the sum of the squared distances between each centroid of a group and the other points of the same group (called inertia) is gradually reduced. The algorithm is not deterministic: the result strongly depends on the choice of the initial points. However, in Python a clever choice is done and, moreover, the algorithm is run multiple times before choosing the choice giving the best results in terms of inertia. Since this usually decreases as the number of groups grows, this parameter is usually chosen by looking for an elbow in the relationship between inertia and number of groups.

- **Hierarchical agglomerative clustering** consists in starting with individual groups, which are gradually merged together until all the records belong to the same group. The ordering in which these mergings are performed depends on the criterion followed for the computation of distances between pairs of clusters. In this case, Ward linkage has been used, which means that the objective function, to be minimized, is the sum of squared differences within all the clusters, very similar to the strategy of k-means. Since there are n possible clusterings, with n being the number of samples, one of them must be chosen. This can be done in two ways: by specifying the threshold distance over which two clusters must not be joined anymore (the minimum computed

distances between two clusters increases as the groups get larger, because the pair which reaches the minimum is joined, so it won't appear as a pair at the following iteration), or by telling the desired number of groups. The latter has been chosen, which is another similarity with k-means.

- **Density-based clustering** consists in forming groups according to density of points in the hyperspace formed by the variables: two parameters are used to set the minimum density required to form a group; all points standing in a lower-density region are classified as noise.

Apart from inertia, which is mainly exploited to choose the optimal number of clusters in k-means, but not for comparisons with other kinds of clustering, other two metrics have been used to evaluate each algorithm:

- The **Davies-Bouldin index** is a measure of the compactness of each cluster, together with the separation from other clusters. Actually, for each cluster, the maximal ratio between the sum of the spatial dimensions of itself and another cluster and the distance between the two clusters is computed. Then, the values are averaged. Therefore, it is better to minimize this metric and the best possible value is 0.

- The **silhouette score** is the average, over all the samples, of a certain coefficient. For this purpose: two quantities must be computed: a is the mean distance from the other elements of the same cluster, while b is the mean distance from the elements of the nearest cluster. Then, the silhouette for a certain point is defined as the ratio between b-a and $max(a, b)$. It ranges from -1, which is the worst case, to 1, which represents the perfect scenario.

The variables considered in the clustering have been those related to the demand, associated to each pair stop+route, occurring in the different time bins. Days have been again split by their type (working, half-holiday, holiday): for each time bin, the total demand and its variance across the days of that type (again, variance must be intended as explained above) have been computed. Therefore, since in some time bins there aren't validations at all, about 20 time bins have been considered, so the variables were about 40. Only after the clustering, some other variables have been added to each sample, in order to describe each group:

- Latitude and longitude of the stop point, to see its position on a geographical map. Since the clustering was instead based on the pair stop+route, the same point on the map may simultaneously be assigned to different groups.

- The kind of place ($TIPO\_LOC$)

- The density of population of the census section

- Number of routes and trips passing by the stop point on that day (as computed by the function *compute_stop_stats*). Since this attribute may depend not only on the stop point but also on the day (some trips or routes may not be ran or change their planning in some days, thus affecting the total number of trips and routes passing by the stop), it has been necessary to compute the mean of its values across the days, for each stop point.
- Whether the stop point was the first stop of the trip (expressed by the variable named *terminal* in the previous Chapter). Also in this case, the mean across the days has been computed. Therefore, a value between 0 and 1 for some stops shouldn't be surprisingly, because at the end it represents the relative frequencies of validations when that stop was the first of the trip, with respect to the total number of validations occurred at that stop.
- The total number of validations, divided by category of user (student, retired, other)

Density-based clustering techniques have turned out to give poor results: depending on the choice of the two parameters, different outcomes have been observed, ranging from very bad values of silhouette and/or Davies-Bouldin index, to very unbalanced groups (indeed, more than those given by the other algorithms), or also with many samples classified as noise. Therefore, the results shown in the following Chapter involve only k-means and agglomerative clustering: these have been compared in terms of Davies-Bouldin index and silhouette, but for k-means also the trend of inertia has been taken into account.

Chapter 7

Results and discussion

7.1 Analysis of the demand

The statistics computed in Section 5.3 have shown that the variables affecting the most the incoming demand are of three types:

- **Time**, especially type of day and hour of day
- **Features of the user and the travel document**, especially category of customer, age and kind of travel document
- **Elements related to the offer** (trips and stop points)

For what concerns time, mainly Figure 7.1, but also all the other ones in which there's at least one temporal dimension, show a great variability of the demand across different days and hours of the day. Specifically, in this Figure all the charts have the timeslot on the horizontal axis: they show the percentage of validations with respect to the total, except for the second chart, where there is the mean for each weekday. The first shows that the timeslot with the greatest total demand is that between 7 a.m. and 8 a.m., as it could have been expected, since many students and workers leave their home during this interval. However, the increasing starts already in the previous hour, due to people who have to make a longer journey. Then, during the following hours of the morning the demand is quite low, but a second peak is observed at lunch time (especially between 1 p.m. and 2 p.m.), when students and part-time workers come back to home. It is instead a surprise, in a certain sense, that there isn't another true peak, but just a slight increasing, in the late afternoon, when many workers leave their offices: the reason of this will be clear when analyzing the demand with respect to the category of customer.

The second and third chart allow to compare the demand across different days and they should be analysed at the same time. They both consider the validations

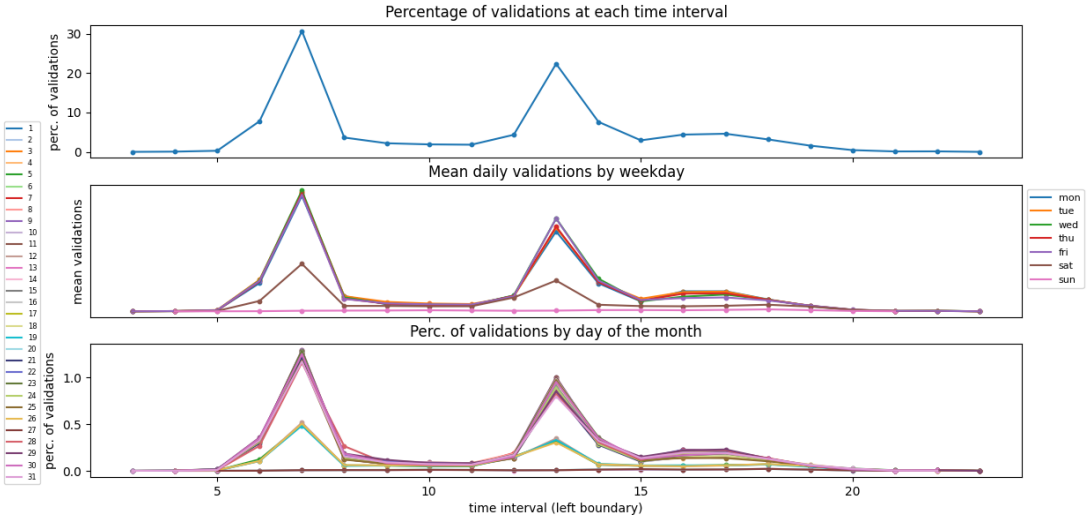

Figure 7.1: Trend of the demand across timeslots, weekdays and single days of the sample month.

during the sample month (October): since in the second chart the seven weekdays are compared and not all the weekdays have the same frequency during a month with 31 days, the mean daily number of validations is computed, instead of the percentage on the total. This allows to realize that in the five working days the total demand is quite constant (if the timestamp is fixed), while on Saturday there is a different behaviour, in the sense that the two peaks are less pronounced, and on Sunday it is very flat and close to zero. This impression is enforced by the third graph, where there are the same three different trends of the demand: even if it is difficult to associate each day of the month to a color, it is very likely that, on each of them, the pattern is the same shown in the previous graph for the corresponding weekday. This justifies the splitting, made in some of the following charts, among working days (from Monday to Friday), half-holidays (Saturdays) and holidays (Sundays); in other words, it will be often unnecessary to consider all the seven weekdays, since there aren't notable differences between the five working days.

Users and travel documents have instead been considered in two possible ways: of course, it is more important observing their distributions based on the validations table, which means that each user or document has weight equal to the number of times it is associated to a validation. However, it is also interesting, mainly for a comparison, to see the distributions of users and documents in their own tables, where each counts as one, not depending on the frequency in the validations table. However, as already said, more than half of the users and travel documents don't appear in the validations of October; therefore, it is better not considering them at all. The partitionings based on the validations table are shown in Figure 7.2. It is

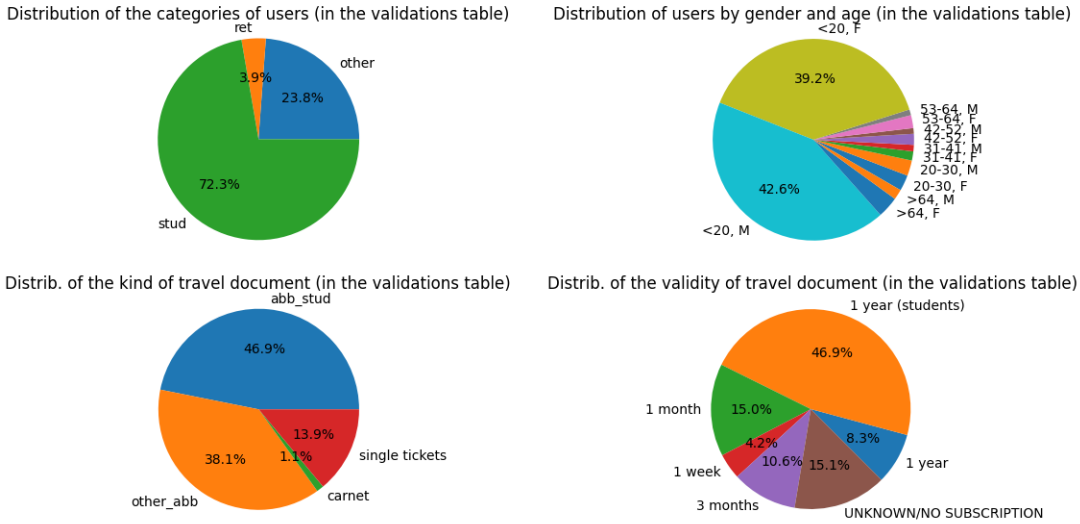

Figure 7.2: Distribution of the users and of the travel documents, based on the validations table.

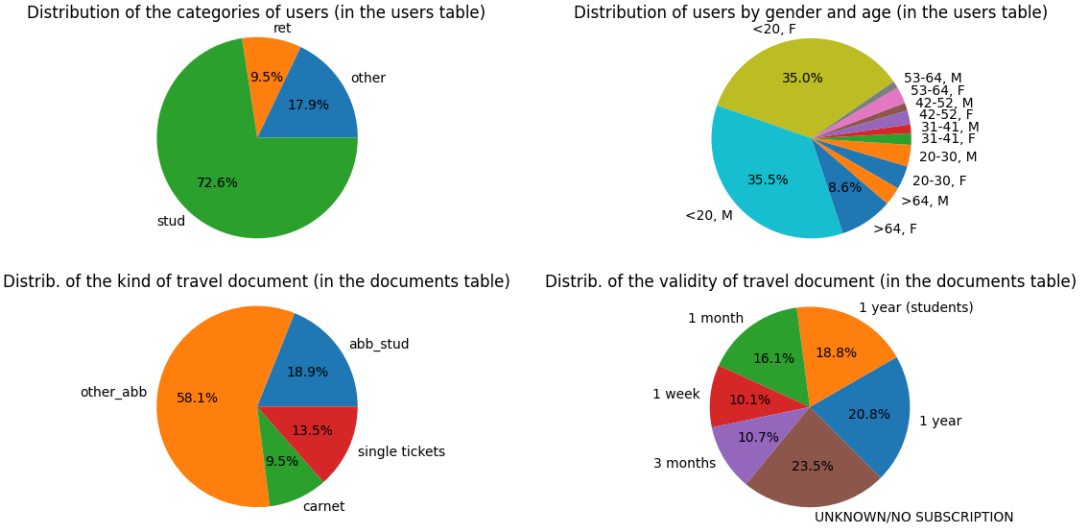

Figure 7.3: Distribution of the users and of the travel documents, based on their own tables.

evident from the two top charts that a big majority of the users are very young and thus many of them are students; this explains also why the total demand in the late afternoon is not very high, with respect to other timeslots. It can also be seen that men are more than women in the low age ranges, while the reverse holds

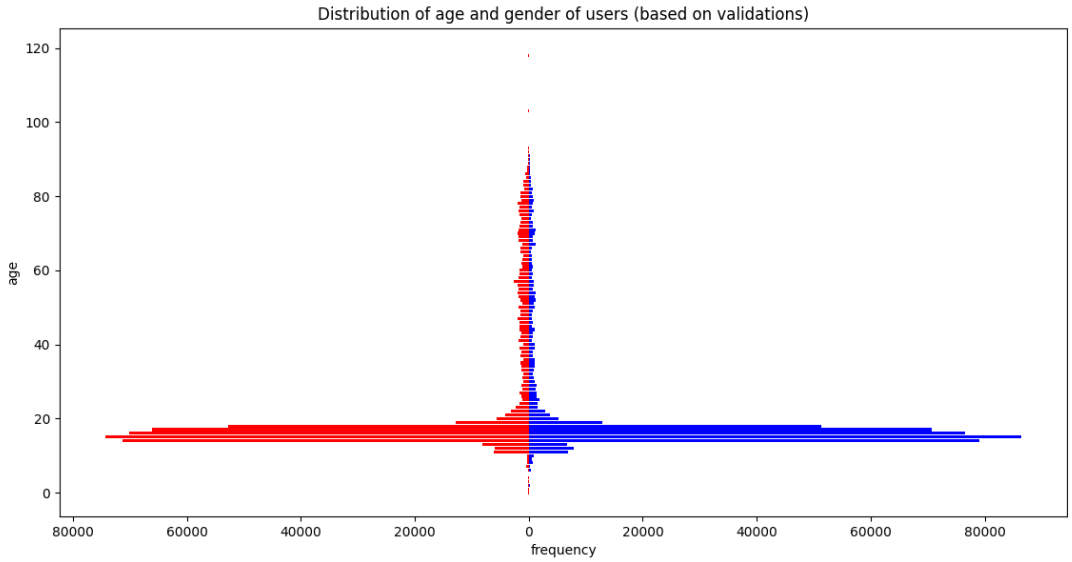

Figure 7.4: Demographic pyramid for the distribution of the customers, based on the validations.

from 31 years on; anyway, the difference is in both cases small. The prevalence of students is mirrored in the two bottom graphs, where the travel documents reserved for them are the majority: now, it is only a relative majority, so probably there are also students with other kinds of travel document.

As said above, the splittings based on the documents and users tables (shown in Figure 7.3) are useful mainly for comparison: in this case, the relative frequency of students, young people, and especially of the related travel documents is smaller than it was in the previous graphs. This means that not only these categories prevail, but also that, in average, they travel more than others.

For a more clear visualization of the distribution of customers according to their features, the demographic pyramid shown in Figure 7.4 is better: it confirms that among young people men (in blue) prevail over women (in red), and viceversa from about 30 years on. It shows also some extreme values which have been already justified: very young children could travel on school buses, while people older than 90 probably don't really travel, but they may lend their document to some younger relatives or friends.

In Figure 7.5, there is the partitioning based also on the category of customer: for a more compact visualization, the frequencies are shown on a logarithmic scale. This allows to see that, as expected, many young users are classified as students and until about 55 years there aren't people recognised as retired (even because too young people in this category have been removed in Section 4.1). It also confirms, as previously anticipated, that customers above a certain age are no more classified

as retired, enforcing the hypothesis that their travel document is used by younger people and, in some way, it is recognised that they are not retired.

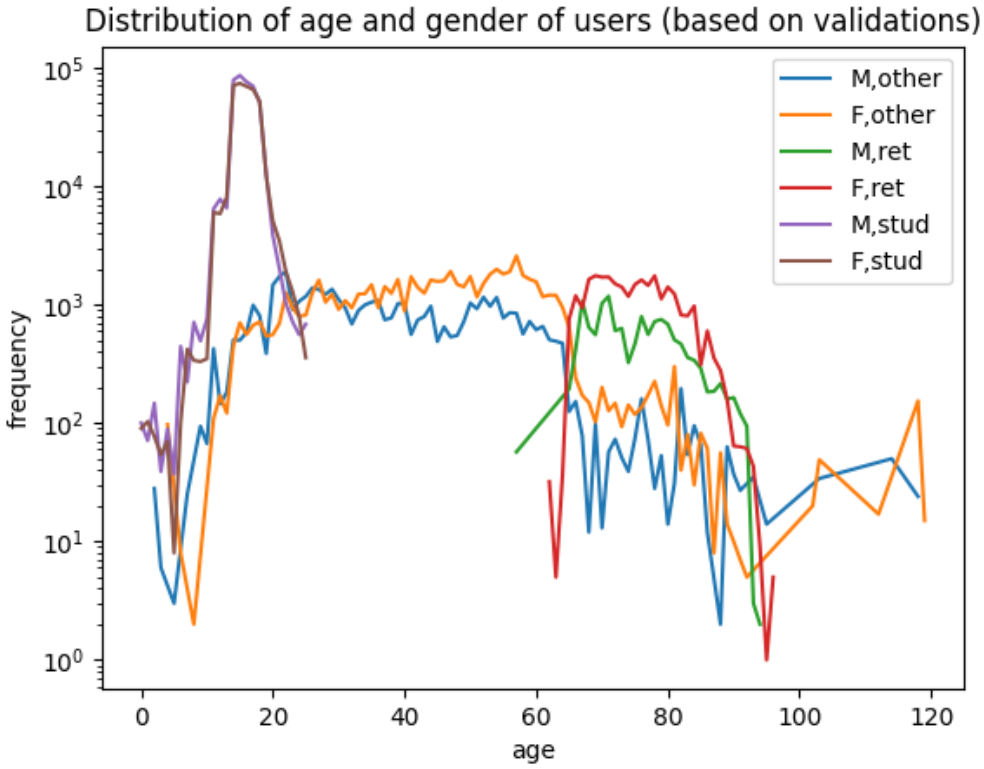

Figure 7.5: Distribution of the customers, based on the validations table.

Then, the focus has been moved on the three found categories of customers, by considering also temporal variables: in Figure 7.6, there is a further split according to the kind of validation (with transport credit or not), in order to realize if there are differences also across this variable. The percentages of validations related to each category of user and type of validation are computed with respect to the total number of records related to that type of day. It can be seen that there are very few validations with transport credit (except for holidays, in which, however, the total number of records is negligible), so this kind of splitting will be abandoned in the following. For what concerns categories, two main patterns are observed:

- The difference between students and other categories is almost fully created during peak hours (from 6 a.m. to 8 a.m. and from 1 p.m. to 3 p.m.), obviously except for holidays, while in the other timeslots also this category shows a low demand. The percentages on the vertical axes of the first two charts must not

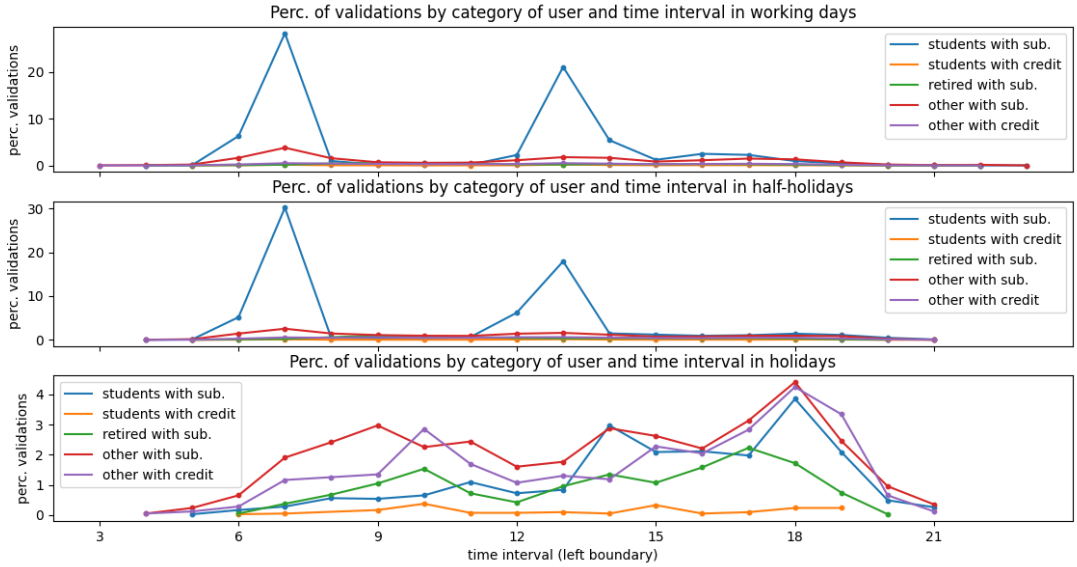

Figure 7.6: Trend of the demand across timeslots, type of day, category of customer and kind of validation.

cheat: they don't mean that on Saturday there are more validations than on weekdays, but just that the relative frequency of students is highest (probably because many workers stay at home, while only few students do the same).

- The trend on Sundays is completely different: on this day, people mainly move for leisure and probably with private means. Thus, students are no more in majority and the highest demand is found during the late afternoon.

For a more complete analysis, the two dimensions have then been reversed (see Figure 7.7): in each graph, there's the trend of the demand along the day for each weekday, for a certain category of customers. Again, since the cardinalities of the weekdays are different, the demand needs to be normalized. For students, all is as expected: the trend is similar to that of the total demand, because they give the highest contribution to it; for *others*, there's a quite high demand also in the late afternoon, a pattern which had remained hidden in the previous charts. For retired people, the trend is completely different: again, as being few customers, it had not been noticed so far. They move mainly during the late morning, but the surprise is that the mean on Tuesdays is notably greater that on other working days and Saturdays. An explanation for this may be some special events (as already said, there aren't information about them) which are repeated with weekly frequency: in this case, every Tuesday. Of course, it must be something that attracts old people, for example a fair in some towns.

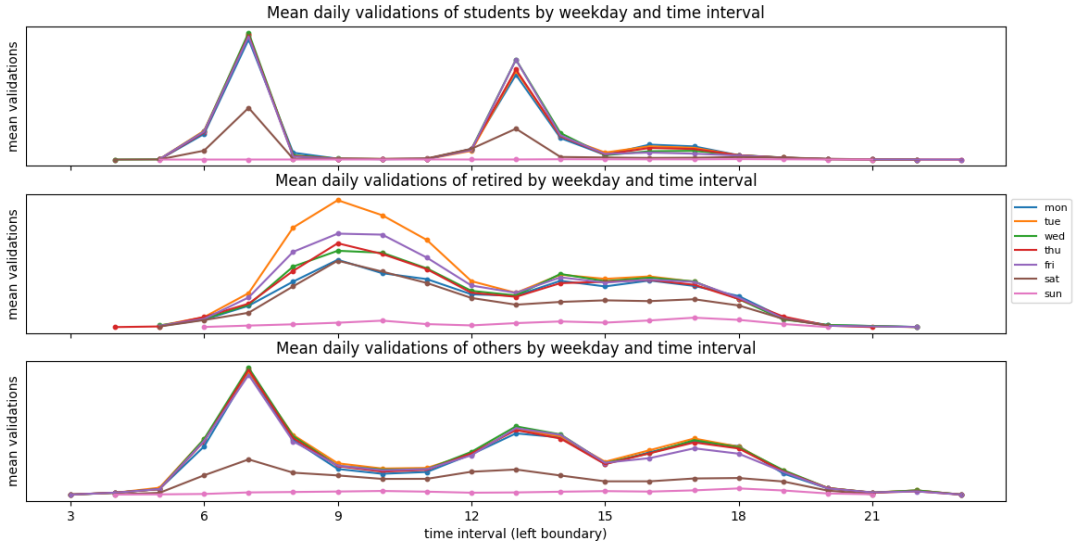

Figure 7.7: Trend of the demand across timeslots, weekdays and categories of customers.

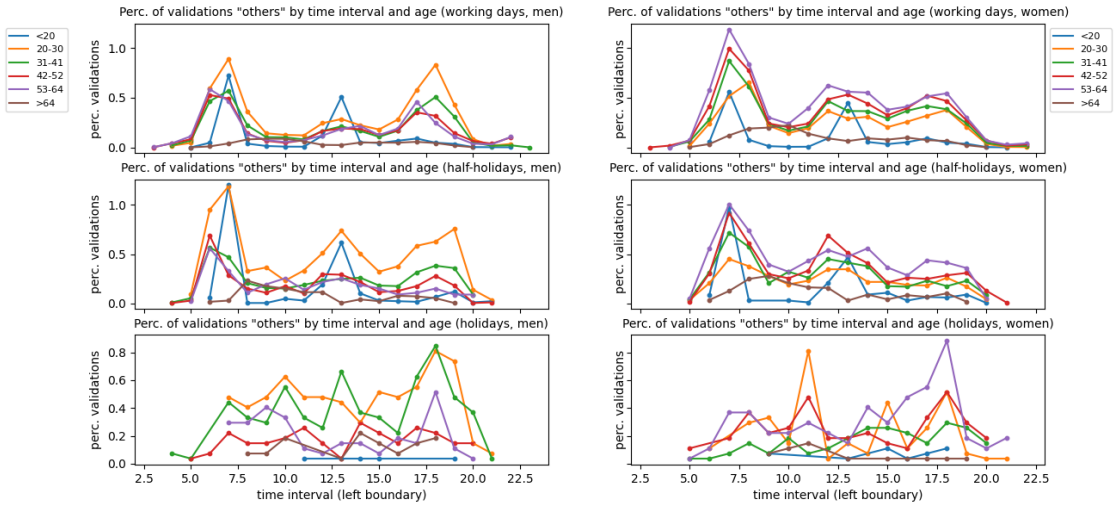

Figure 7.8: Trend of the demand across timeslots, type of day, age range and gender of *other* customers.

Figure 7.8 is useful for a deeper analysis of *other* customers: so far, they have been considered all together, except for the previous pie charts in which, however, there wasn't any temporal dimension. Now, they are split according to age and gender (men on the left column, women on the right), while the temporal attributes

are, as usual, timeslot and type of day. The percentages of validations for a certain type of day are computed with respect to the total number for that type of day. From Monday to Saturday, among men, the customers with the highest relative frequency seem to be those up to 41 years old. Among women, instead, the prevailing age range goes from 31 to 64 and, as already observed, their relative frequency is in general slightly higher than for men. However, the six age ranges shown for each gender have few similarities one with another across the timeslots and, again, on holidays there's a great difference with other types of days.

Figure 7.9: Trend of the demand across timeslots, type of day and kind of travel document.

Also the features of the travel documents are analysed together with the temporal attributes: in Figure 7.9, each chart shows the usage of each kind of document during a certain type of day. Again, each percentage is with respect to the total for the same type of day. This mirrors the distribution shown in the pie charts above: as expected, on working days and half-holidays, subscriptions for students prevail, especially in the two peak timeslots; at the second place, there are other kinds of subscriptions, which are likely to be used by workers, another category which regularly travels, so it finds convenient this solution, rather than tickets, which actually are poorly used during the whole day. The situation is completely different on holidays, probably because travels are occasional for many customers. Therefore, the most used kind of document is the single ticket, but, of course, if the subscriptions cover also the holidays, they are however used if necessary, so they don't disappear.

The last big dimension of analysis is represented by stop points and trips: in this

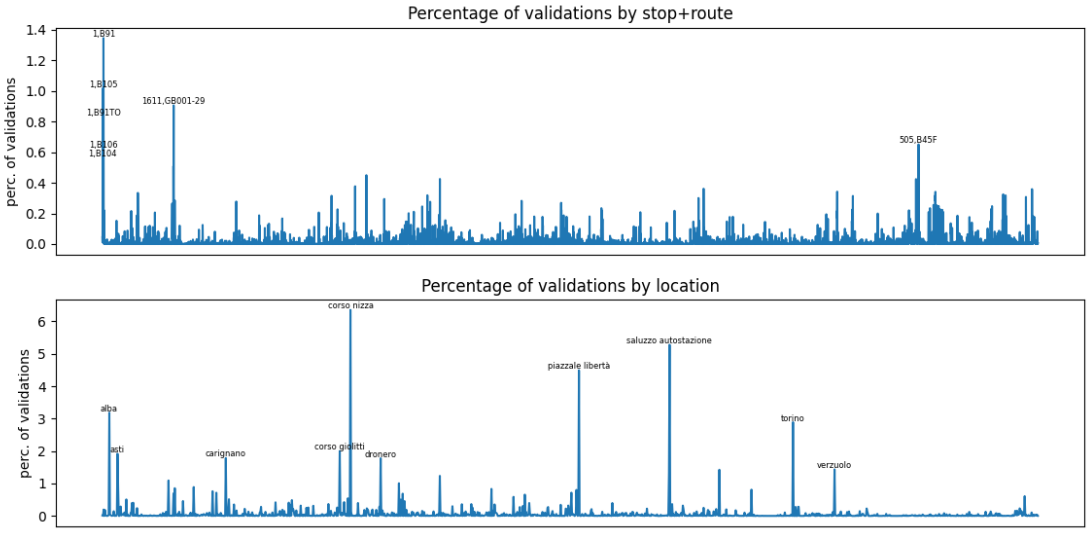

Figure 7.10: Trend of the demand across stop points and locations.

case, the classification is not trivial and, for the moment, each item is individually considered. First of all, in Figure 7.10 stop points and locations are taken into account. In the first case, validations without user code are excluded, because the stop point may be inexact. Since many stop points are met by multiple routes, each stop is separately taken according to the incoming demand for each route. In the first chart, the pairs stop+route with the highest number of validations are labelled with their codes, separated by comma: each percentage is computed with respect to the total during the sample month. It is immediately evident that the most crowded pairs are related to stop number 1, which corresponds to the train station of Saluzzo, near Cuneo, followed by numbers 1611 (located in an important street of Cuneo) and 505 (the train station of Alba, again near Cuneo). As already observed, it is an area rich of stop points and validations. For what concerns locations, they must be considered carefully, because, as being extracted by the stop name, sometimes they don't explicitly tell the town they belong to, but they are useful for trying to geographically group the stop points. The most frequent ones are again in Cuneo (*corso nizza* and *piazzale liberta* are two central places), Saluzzo and Alba. From the percentages, it can be estimated that the 7 stop points with the highest number of validations cover more than 6% of the demand, while in the first three locations occurs about one validation out of six.

For what concerns trips, from Figure 7.11 the distribution of validations looks slightly more balanced: just few of them cover more than 0.2% of the total demand: actually, each percentage is with respect to the total of the sample month, after having excluded records without the trip code. The second and third chart show

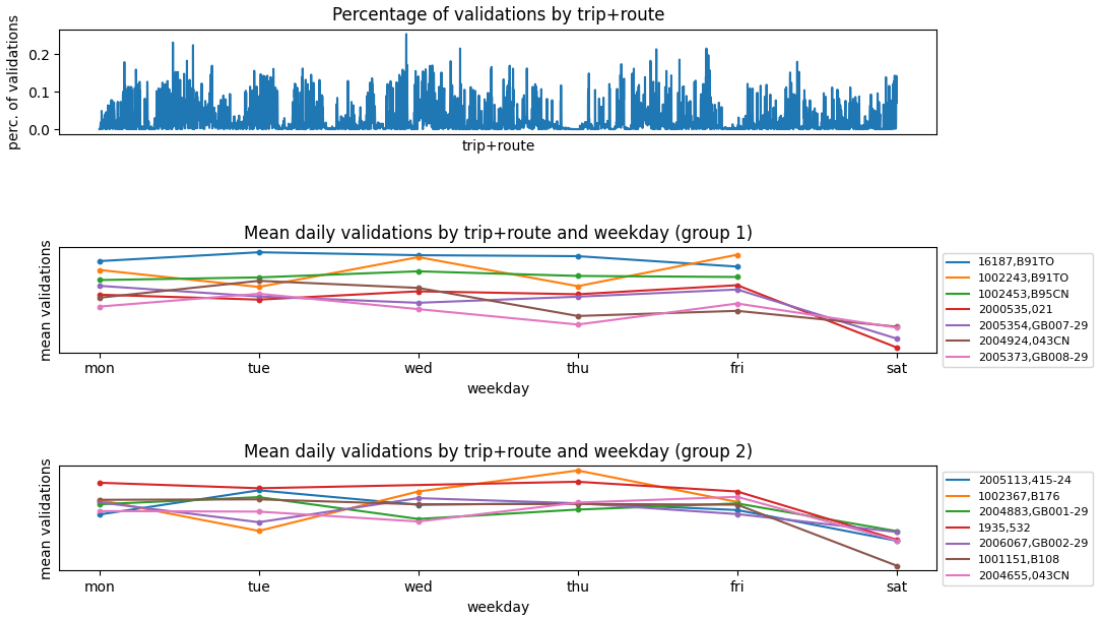

Figure 7.11: Trend of the demand across trips.

the trend of the related validations for the most recurrent trips (from the first to the seventh and then from the 8th to the 14th): as for stop points, also the route is shown (in this case, the route doesn't affect the partitioning, since each trip is unique even if the route is not given). Observing the demand of a certain trip across the daily timeslots wouldn't make sense, since each trip is ran at most once per day. Thus, on the horizontal axes, there are the weekdays: the total absence of Sundays and holidays means that these trips, even if they are the most crowded, aren't ran on these days (or, very unlikely, that they don't have passengers at all). Some of them are missing also on Saturdays, when in general almost all the trips show an evident fall of the demand. Many of these trips have a quite flat demand across the five working days; other ones are more oscillating. The most recurrent trips belongs to the route B91TO (which is very important, because it joins Saluzzo and Torino), but there are also many trips starting from or arriving to Cuneo, for example those related to routes B95CN, 021 and all the codes starting with GB.

In order to investigate the variability of the demand at stop point (plus route), customer or trip level, the boxplots shown in Figure 7.12 are a good tool: since there are thousands of items, it's impossible to show all of them, so only the first 14 elements in each ranking are displayed. The boxplots are sorted by decreasing total demand; each of them shows the distribution of the daily demand, extending from the lower to the upper quartile values of the data, with a green line at the median; the whiskers extend from the box to show the range of the data. Flier points are

those past the end of the whiskers. Part of variability may be due to the lack of any partitioning based on the type of day. The stop point 1 seems to have the highest variability, especially for route B91 (which goes from Saluzzo to Cuneo). Among the most recurrent customers, it's evident the absence of "impersonal" users, whose code starts with 2, which means that the related cleaning phase has worked well. Also in this case, there are strong differences in the variability of the demand. For what concerns trips, the variability looks slightly smaller, except for codes 1002243 and 1035.

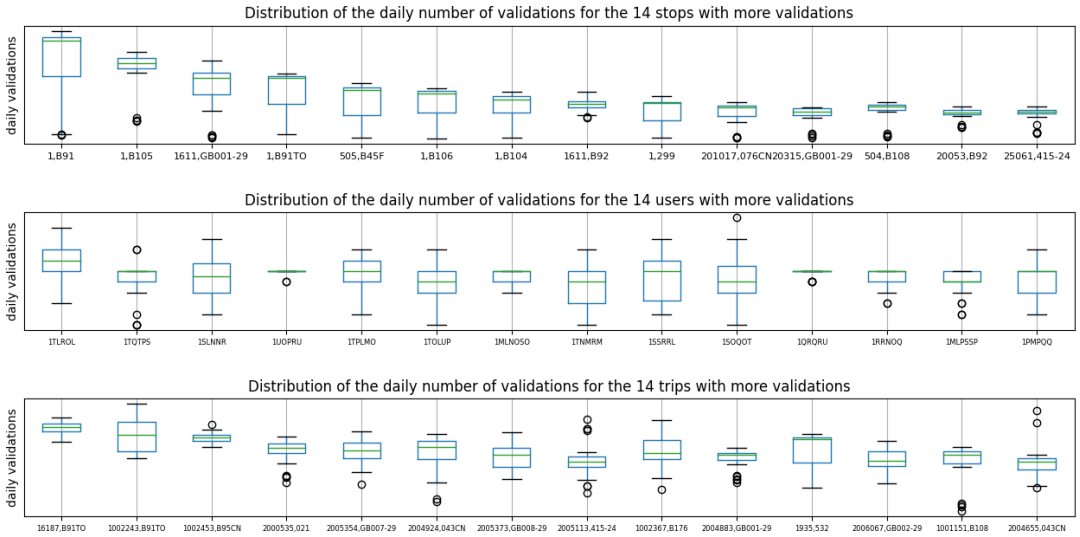

Figure 7.12: Boxplots with the distribution of the daily number of validations for the 14 stops, users and trips associated to the highest total number of records, respectively.

The very last dimension of analysis is the weather condition: as already said, this has been recorded on a daily basis, for each municipality, therefore it may be not very precise, especially in temporal terms. Figure 7.13 shows in each chart, for a certain type of day, the mean daily number of validations for each weather situation (recorded at that stop point on that day). If for some labels there aren't values at a certain timeslot, it means that during the whole period there wasn't any validation at any stop point with that weather condition on that type of day (this is likely to happen if that condition was recorded for few pairs day+stop point). Of course, even if the mean demand is computed, this representation is strongly influenced by the frequency of each weather situation: it can be seen that some of them are completely or partially absent, especially those which are typically extemporaneous, so they are unlikely to be associated to a whole day (such as snow or fog). In a similar fashion, also due to the period of the year, it has never been

recorded, in any place, a fully sunny day. However, an expected pattern can be noticed, on each type of day: bad weather (especially rainy and stormy) seems to enhance people to use public transport. This may be due to the fact that, when the weather is bad, there are more people switching from walking or going by bike to taking a bus, rather than those who usually use public transport but in bad days prefer to go with their own cars: maybe, some of them don't have a car at all.

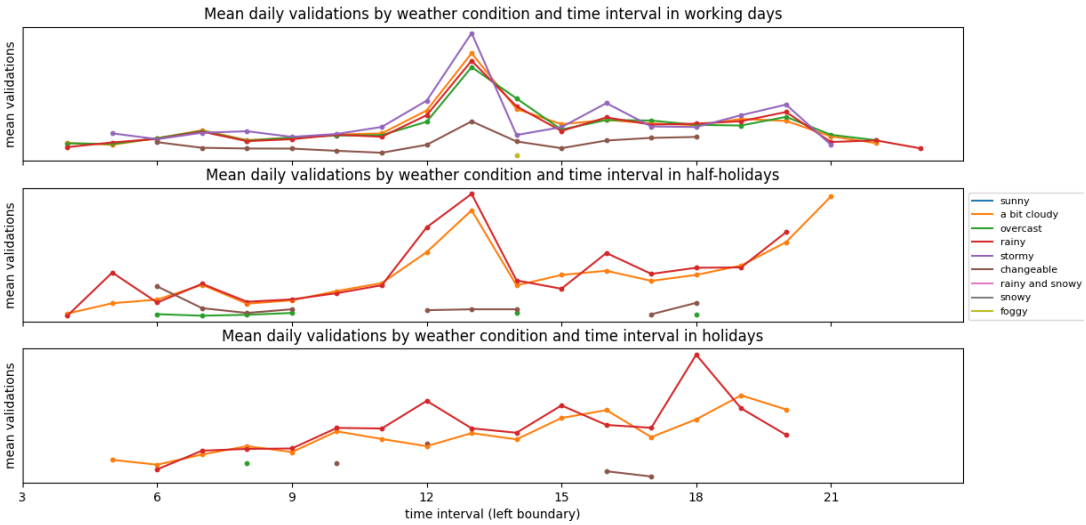

Figure 7.13: Trend of the demand across timeslots, type of day and weather condition.

7.2 Characterization of the stop points

As already said, density-based clustering has given worse results with respect to agglomerative and k-means: therefore, for each type of day, the best algorithm has been chosen between these, by plotting the trend of the quality metrics for the number of clusters ranging from 2 to 10; also greater values have been tested, but only in a post-processing phase, since in general a too large number of groups makes the results less interpretable. For the selected combination of technique and number of groups, some charts have been plotted. Specifically, for each group, the distribution of its elements across many of the variables added after the clustering: geography, features of the stop point (or of the related census section) and demand divided by category of user are shown.

7.2.1 Working days

For what concerns demand during working days, 6714 pairs stop+route have been considered.

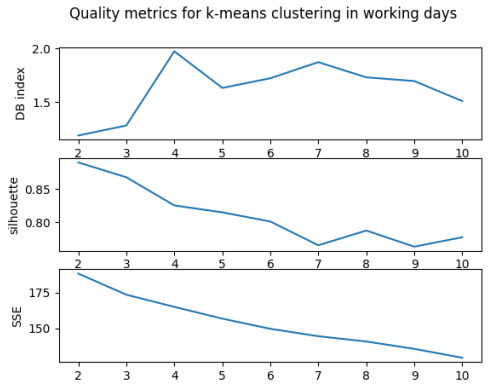

Figure 7.14: Distribution of the quality metrics of k-means clustering across the number of groups on working days.

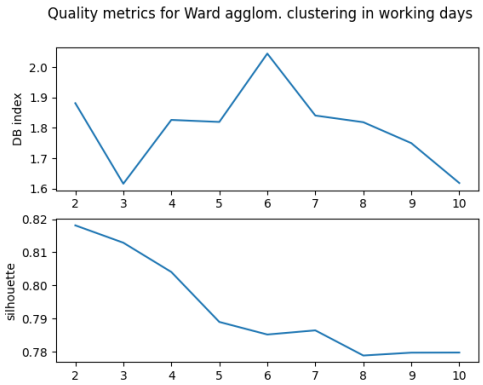

Figure 7.15: Distribution of the quality metrics of Ward agglomerative clustering across the number of groups on working days.

Figure 7.14 shows a quite constant decrease of inertia, except between 2 and 3 groups (in the figures, it will be called SSE, as being the acronym of *sum of squared errors*). Therefore, it is better looking also at the other two charts and to compare them with those related to agglomerative clustering, shown in Figure 7.15. It seems that k-means is preferable, since silhouette is in general slightly higher and DB index is often lower, compared with the same number of groups obtained with agglomerative clustering. The choice of the number of groups looks hard: however, a very little slowdown of the decreasing of SSE can be observed at 6 and 7. Since the second shows a local maximum of DB index and a local minimum of silhouette, the first option is preferred.

For each of the formed groups, the variables listed in the previous Chapter, with the exception of *TIPO_LOC* (which has turned out to be not very characterising) and *num_routes* (which is expected to be correlated with *num_trips*) have been added. Figures 7.16, 7.17, 7.18, 7.19, 7.20 and 7.21, show the geographical position of the stop points of each group: they can be distinguished thanks to the blue marker, which in some cases is bigger just to make them more visible.

Instead, the distribution of the other variables is shown, separately for each group, in Figures 7.22 and 7.23. The histograms describing the same variables (those on the same column) all have the same horizontal scale (even if the numbers

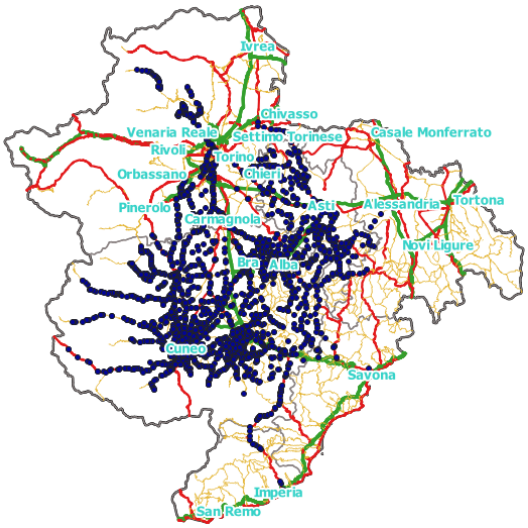

Figure 7.16: Geographical distribution of the stop points of group 0 on working days.

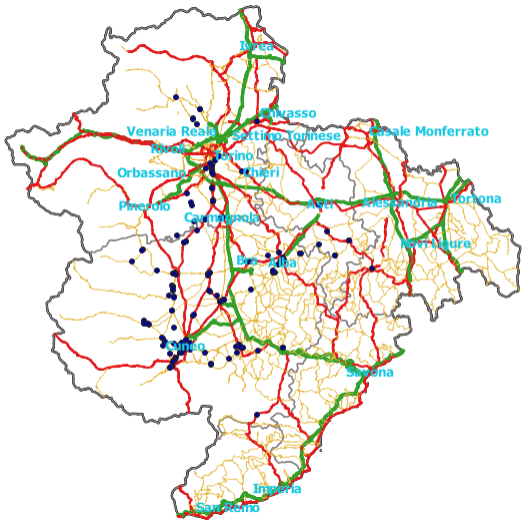

Figure 7.17: Geographical distribution of the stop points of group 1 on working days.

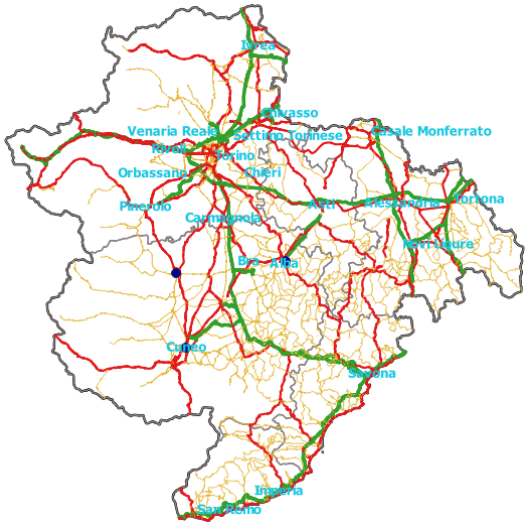

Figure 7.18: Geographical distribution of the stop points of group 2 on working days.

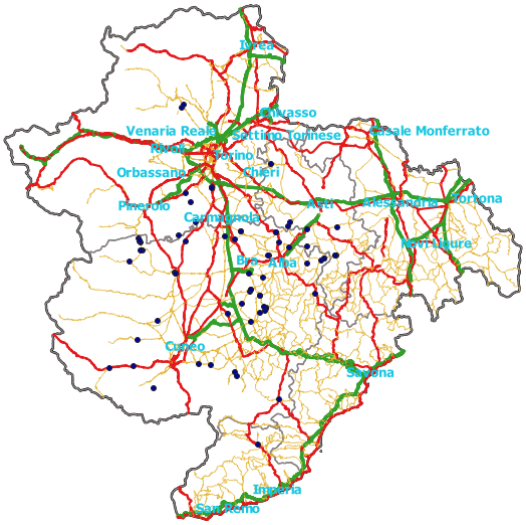

Figure 7.19: Geographical distribution of the stop points of group 3 on working days.

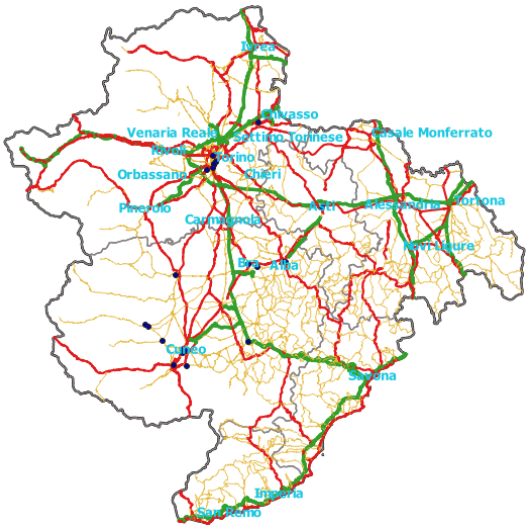

Figure 7.20: Geographical distribution of the stop points of group 4 on working days.

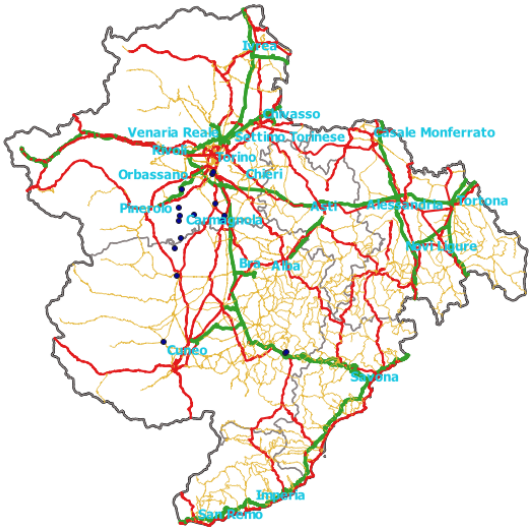

Figure 7.21: Geographical distribution of the stop points of group 5 on working days.

are not always reported), such that the groups can be more easily compared. It seems that the most characterising variables are the number of trips passing for the stop and the total number of validations, especially from students and other people. Group 0 is the biggest (6303 samples), but almost all of its elements are pairs stop+route with a low importance: actually, there are very few validations from all the categories, a very low number of interchanges and they are very rarely the first stop point of the trip. Moreover, this is the group with the highest average of *TIPO_LOC*, which, due to the coding of each value, means that the group includes many stops located in isolated places. The opposite holds for group 2: it contains only 5 samples, belonging to three different stop points which are very important (the train stations of Saluzzo and Alba and another in the center of Cuneo). Actually, its element show very high numbers of validations, especially from students and others, which are the most likely to commute at the train stations. Moreover, they have many interchanges and almost always they are the starting point of the trips (the bar is very thin because all the values are very close to 1). Also in groups 4, formed by 50 elements, there are stop points very important from the offer side (in terms of *num_trips* and *terminal*), but with a lower demand if compared with group 2. However, they are all associated to the first type of locality (the most populated), actually many stops stand around Cuneo and Turin. A similar reasoning can be done for group 1 (high importance in terms of offer,

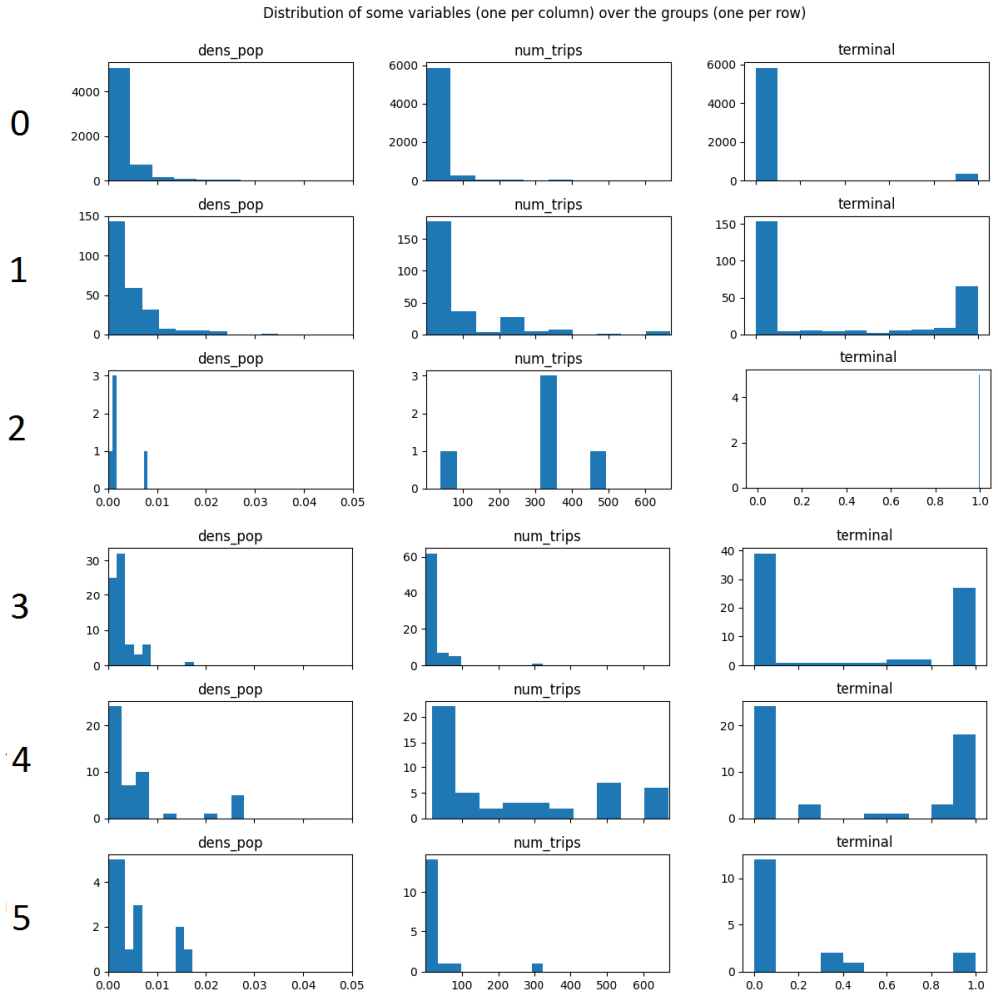

Figure 7.22: Distribution of some variables across the groups on working days.

less in terms of demand), which includes 264 samples more spread in the land. Finally, the demand and the importance of the stops are reduced for group 3 (75 elements, many of which seem to be out of the towns) and even more for group 5 (17 elements), whose feature is having more or less the same incoming demand from students and other people, differently from what observed so far in this Chapter.

Since group 0 is very big in comparison with the other ones (nearly 94% of the samples), a further partitioning has been tried on itself. However, as it can be inferred from the dendrogram in Figure 7.24, it has been very difficult to split that group. Starting from the top, each bifurcation of the tree represents the separation of a group into two children and the distance (from the root) where it occurs makes it understand how much the two new groups are similar to each other. The different

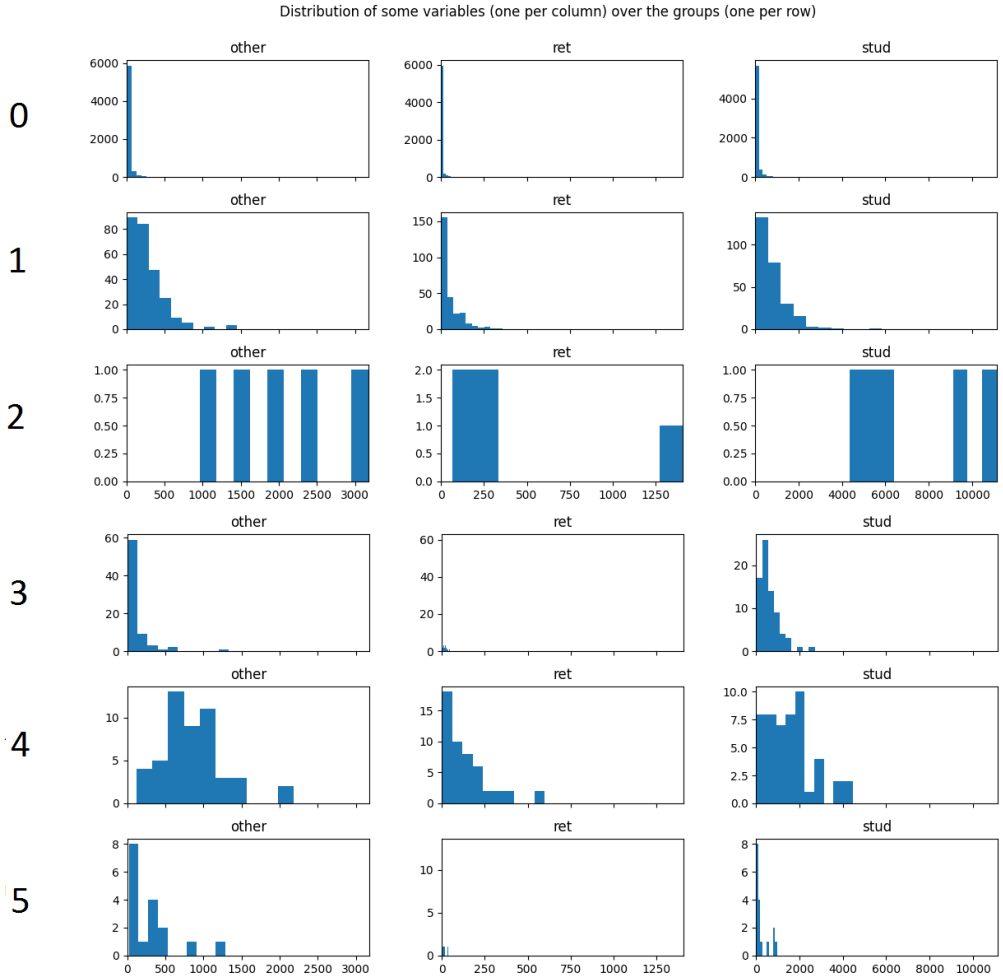

Figure 7.23: Distribution of some variables across the groups on working days.

colors give an idea about a possible partitioning: they show that, if three groups are formed starting from cluster 0 (red, green and yellow), they are in turn very unbalanced. Also with a bigger number of new groups, the situation has proved to be similar, because in the middle of the tree in can be seen that new groups are formed very late, with respect to further partitionings of the green and yellow branches. In the reality, agglomerative clustering begins from the bottom of the tree and joins two clusters at once, but starting from the top is probably more readable.

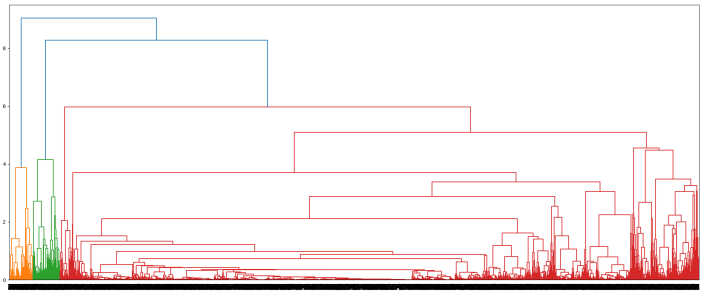

Figure 7.24: Dendrogram related to agglomerative clustering applied on group 0 for working days.

7.2.2 Half-holidays

Figures 7.25 and 7.26 show the trend of the quality metrics of k-means and Ward agglomerative clustering related to the demand, for each pair stop+route, during half-holidays, which during October 2019 are restricted to Saturdays. In this case, 3426 samples have been involved. Again, the decrease of inertia is quite linear; now, for a medium number of groups, Ward clustering looks slightly better: between 5 and 7, the silhouette stays high and nearly constant, while Davies-Bouldin index shows a local decreasing; in this case, the best choice of the number of groups seems to be 7.

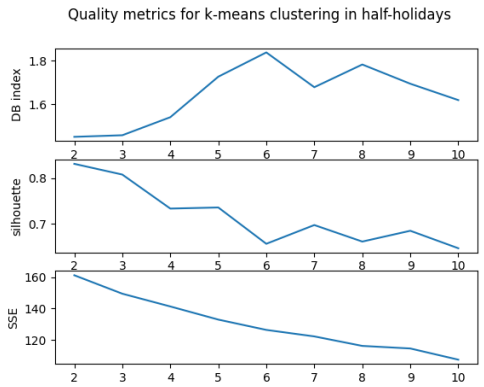

Figure 7.25: Distribution of the quality metrics of k-means clustering across the number of groups on half-holidays.

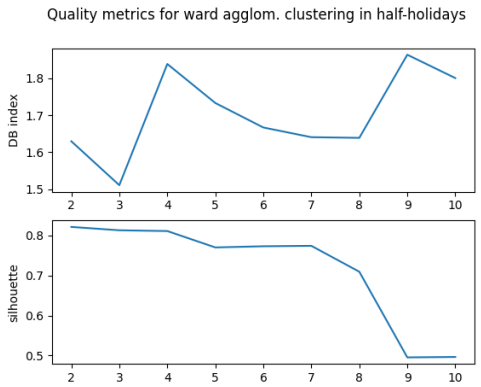

Figure 7.26: Distribution of the quality metrics of Ward agglomerative clustering across the number of groups on half-holidays.

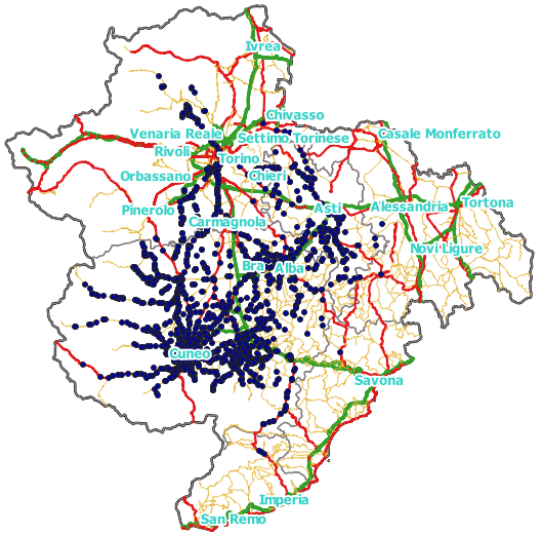

Figure 7.27: Geographical distribution of the stop points of group 0 on half-holidays.

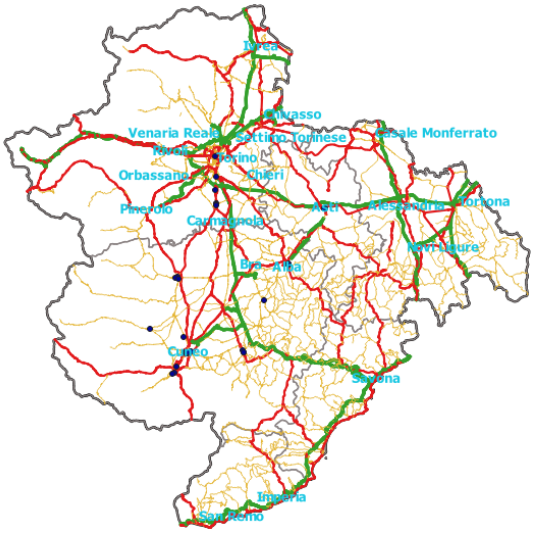

Figure 7.28: Geographical distribution of the stop points of group 1 on half-holidays.

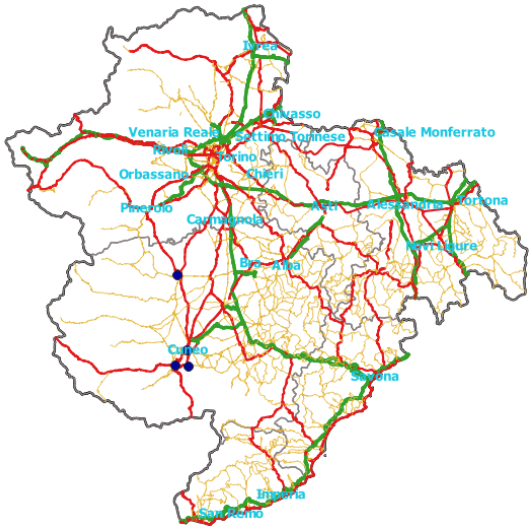

Figure 7.29: Geographical distribution of the stop points of group 2 on half-holidays.

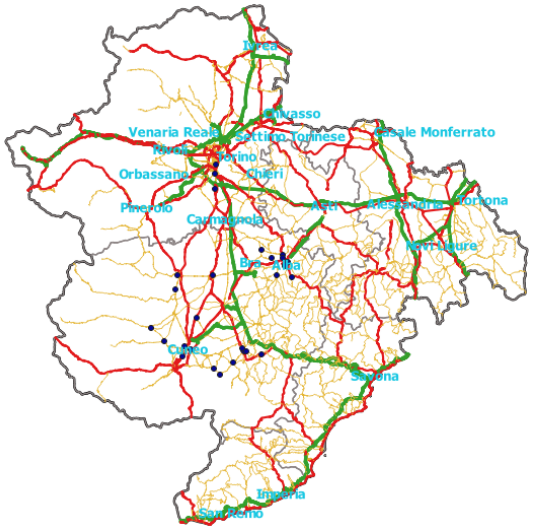

Figure 7.30: Geographical distribution of the stop points of group 3 on half-holidays.

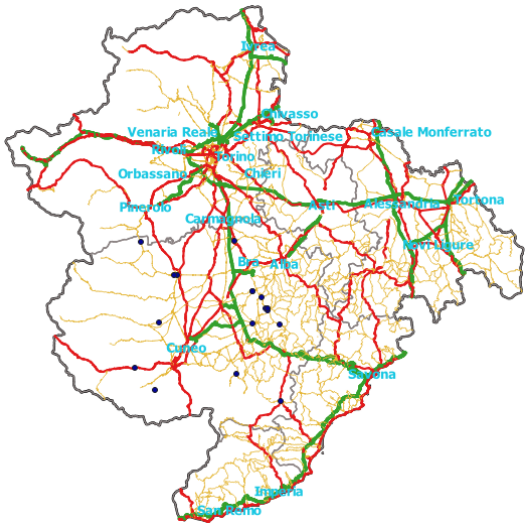

Figure 7.31: Geographical distribution of the stop points of group 4 on half-holidays.

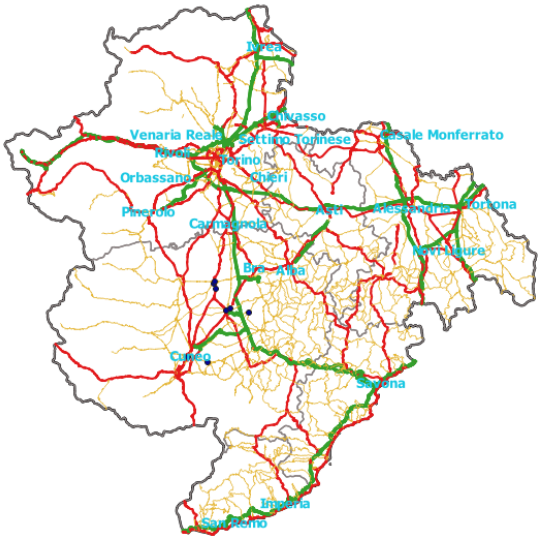

Figure 7.32: Geographical distribution of the stop points of group 5 on half-holidays.

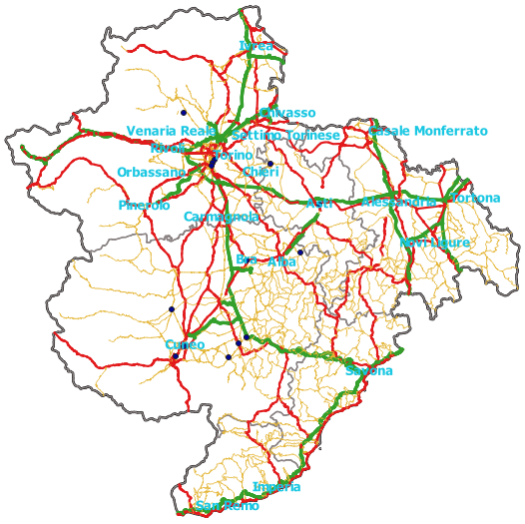

Figure 7.33: Geographical distribution of the stop points of group 6 on half-holidays.

Figures 7.27, 7.28, 7.29, 7.30, 7.31, 7.32 and 7.33 show the geographical distribution of the seven groups, while Figures 7.34 and 7.35 display the distribution of other variables across the groups.

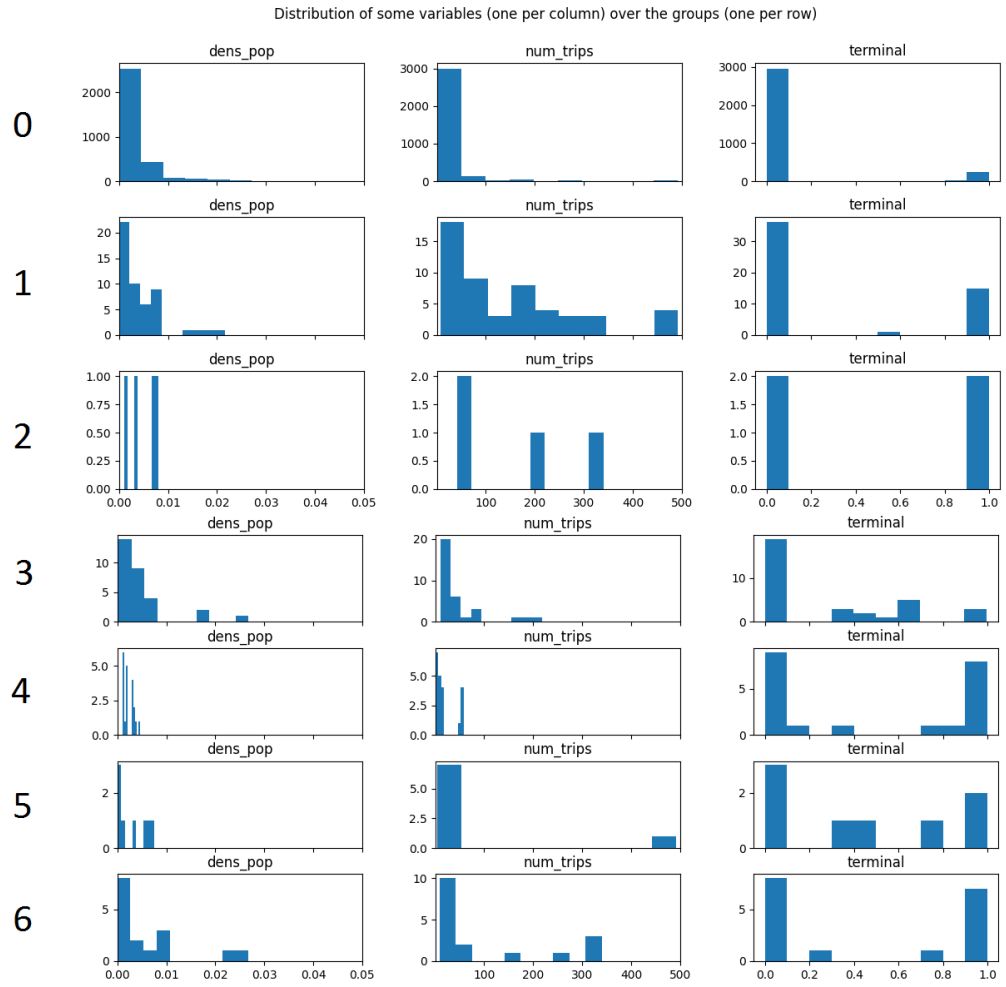

Figure 7.34: Distribution of some variables across the groups on half-holidays.

Now, the most characterising variable seems to be the demand coming from students, followed by that from others and *num_trips*. Again, group 0 is the biggest, as having 3291 samples, but almost all of them have low importance both in terms of features of the stop and number of incoming validations. On the contrary, the 4 samples of group 2 (one is again the train station of Saluzzo, the others are in Cuneo or around) are often crowded stops with many interchanges. Group 3 (33 elements) is formed by stops crowded on average and with low importance, even if near the most important towns, while many of the 52 elements of group 1 are characterised by a great number of interchanges and students. Finally, the last three groups are similar to each other: their elements often have a low value of *num_trips*, very few validations from retired people and, in almost half of the cases,

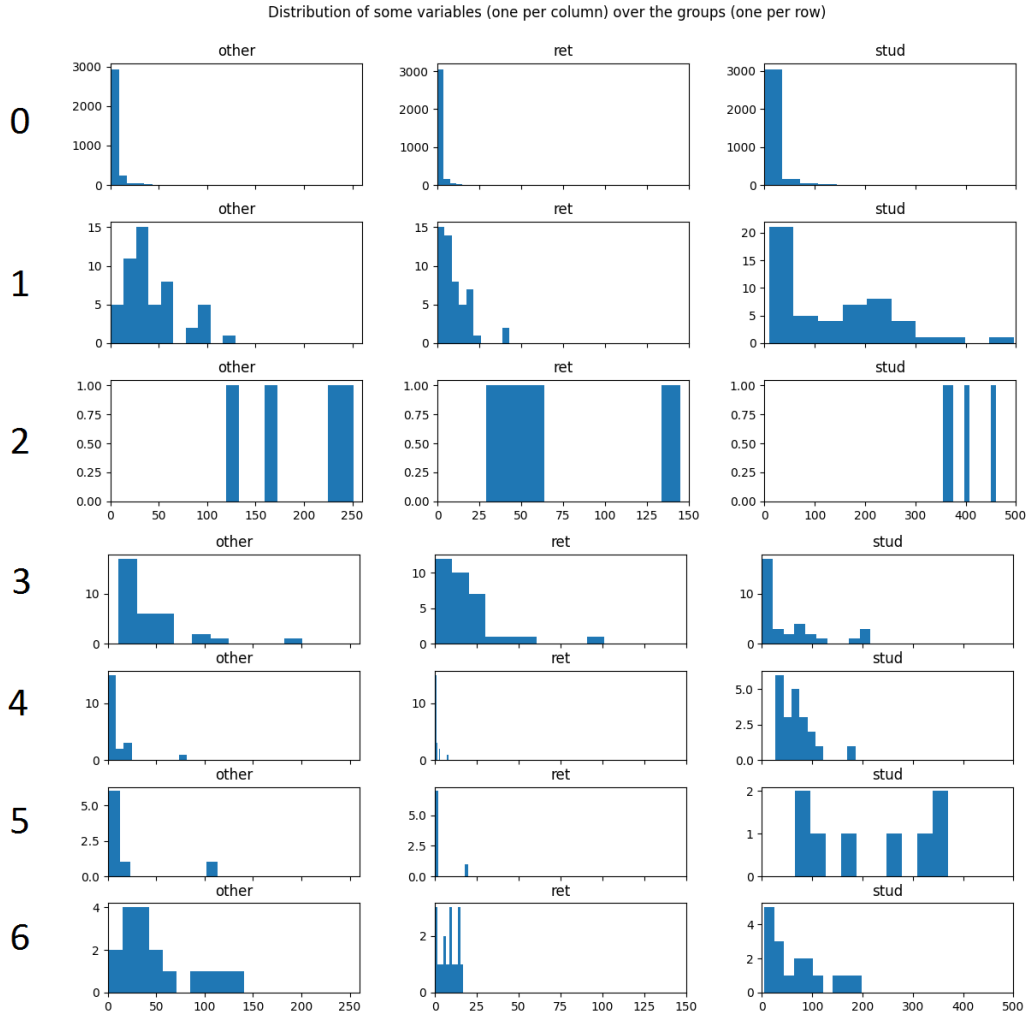

Figure 7.35: Distribution of some variables across the groups on half-holidays.

they are the first stop of the trip. However, there are also some differences: the 21 elements of group 4 show low population densities (as it can be imagined from their positions). Group 5 (8 elements) is the only with kinds of locality different from the first (together with group 0, but this is very large so it's less interesting) and there are especially students. On the other hand, in the 17 members of group 6, the distribution of categories of customers is more balanced between students

and other people.

7.2.3 Holidays

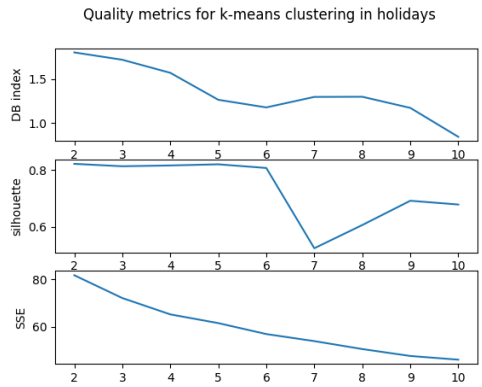

Figure 7.36: Distribution of the quality metrics of k-means clustering across the number of groups on holidays.

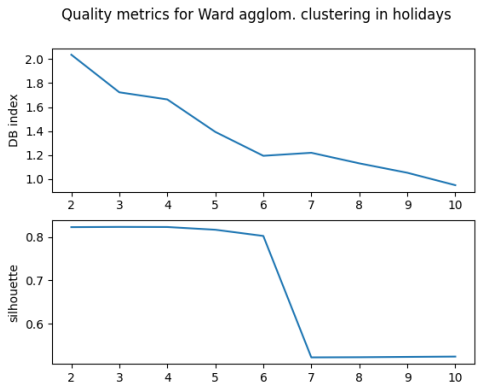

Figure 7.37: Distribution of the quality metrics of Ward agglomerative clustering across the number of groups on holidays.

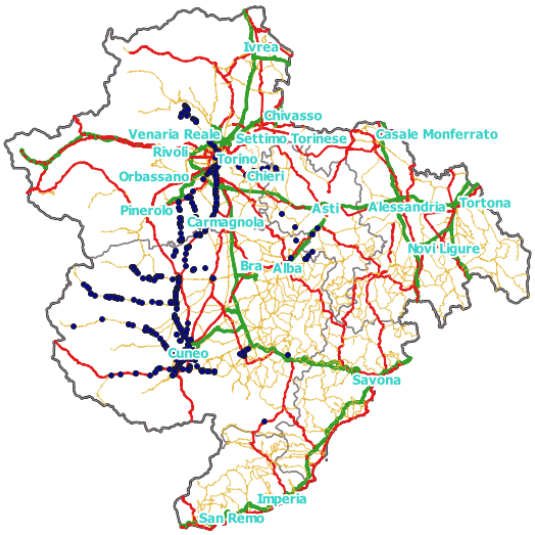

Figure 7.38: Geographical distribution of the stop points of group 0 on holidays.

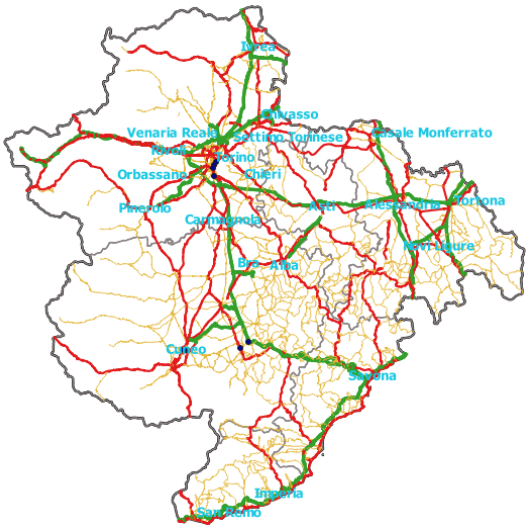

Figure 7.39: Geographical distribution of the stop points of group 1 on holidays.

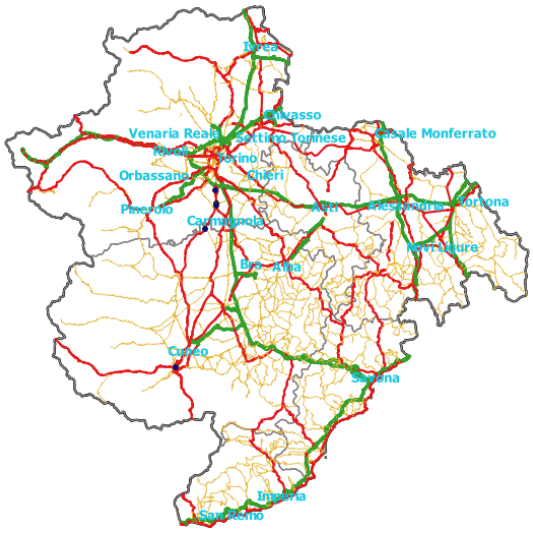

Figure 7.40: Geographical distribution of the stop points of group 2 on holidays.

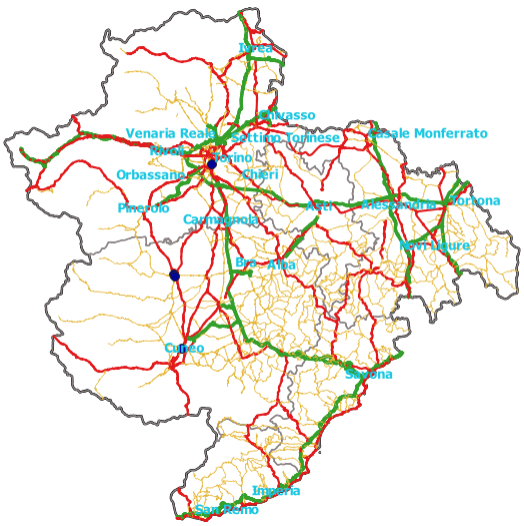

Figure 7.41: Geographical distribution of the stop points of group 3 on holidays.

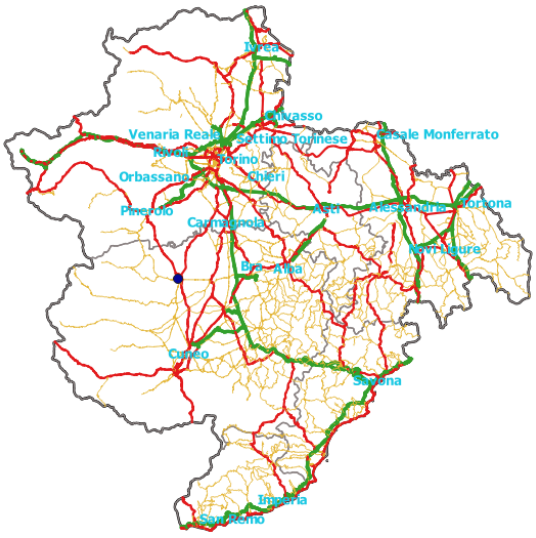

Figure 7.42: Geographical distribution of the stop points of group 4 on holidays.

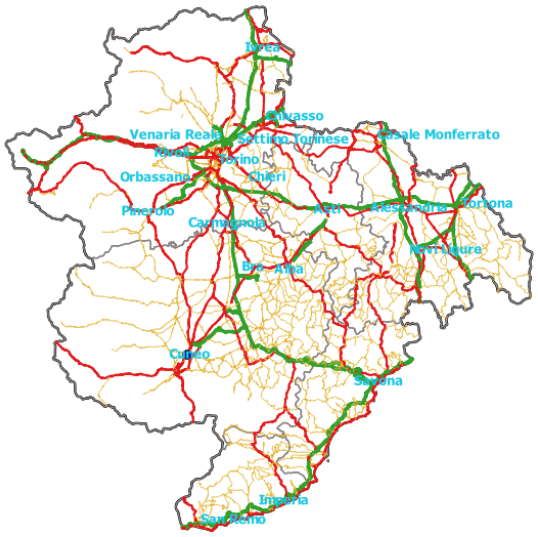

Figure 7.43: Geographical distribution of the stop points of group 5 on holidays.

As already observed in the previous Chapters, the demand on holidays is strongly reduced: actually, there are only 578 active pairs stop+route. By looking at Figures 7.36 and 7.37, the decreasing of SSE for k-means shows at least an elbow (at the

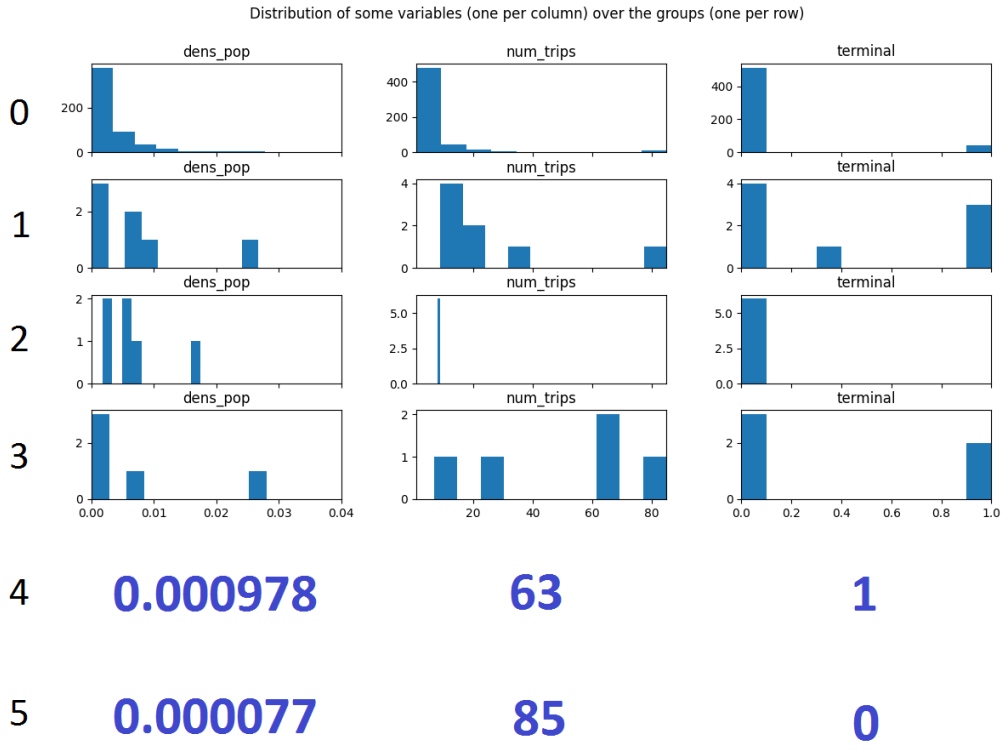

Figure 7.44: Distribution of some variables across the groups on holidays.

value of 4 groups), but at the end the agglomerative technique has been preferred, with 6 groups, where the silhouette is still high, while Davies-Bouldin index shows a local minimum. However, at this point, the quality metrics of the two algorithms have very close values.

Again, Figures 7.38, 7.39, 7.40, 7.41, 7.42 and 7.43 show the geographical distribution of the six groups, while Figures 7.44 and 7.45 display the distribution of other variables across the groups: for groups formed by only one element, the value of each variable is explicitly written.

On holidays, all the stop points located in isolated areas are completely inactive: indeed, all the samples are associated to *TIPO_LOC* number 1. As expected, the demand coming from *others*, together with the number of trips, is more characterising that that from students. Also in this case, the largest group, formed by 557 elements, contains stop points where there are few validations and interchanges and which are very rarely the starting point of the trip. On the contrary, in group 1 (8 samples) there are more important stops, both in terms of features and validations, from all the categories. There are two standing out groups, as being formed by only one element: they are located in the train stations of Saluzzo (the beginning of the route leading to Turin) and Cuneo, respectively. As expected from their position,

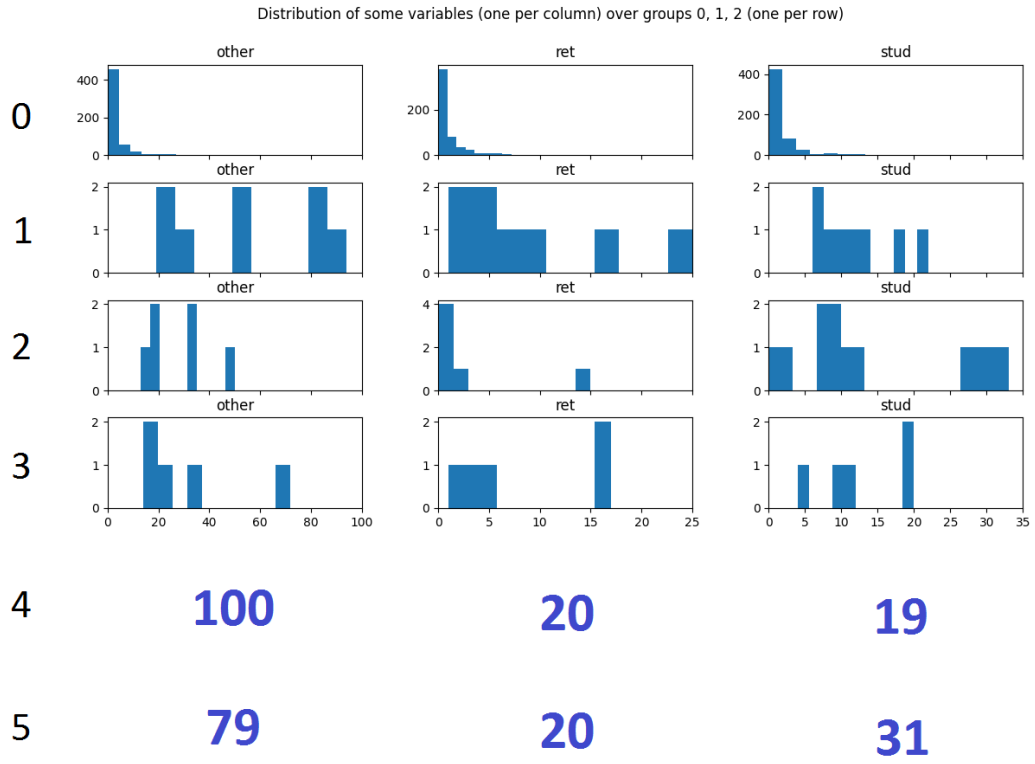

Figure 7.45: Distribution of some variables across the groups on holidays.

they both have many validations and interchanges. In group 3 (5 elements) there is still a good number of validations and especially of *num_trips*; group 2, formed by 6 samples, is similar to the previous in terms of demand, while it's the last one on the offer side.

Chapter 8

Conclusions

8.1 Main results

This thesis has at first tried to observe the distribution of the demand of public transport offered by a consortium located in North-Western Italy, with respect to the indicators which were supposed to affect it the most. The main source for the research has been represented by smart card data, consisting in the validations occurring during a whole month on the buses, most of them associated to a travel document and a customer, for which additional information were usually available. After having properly cleaned the data from several anomalies, the analysis has been focused on several kind of variables:

- **Time.** The behaviour of the customers has been found to be very similar across the working days, with two peak hours at early morning and lunch time, respectively. On Saturdays, the two peaks were less important, while on holidays the trend was completely different and the total demand was notably smaller.
- **Customer.** A first splitting into students, retired and other people has highlighted that the first category represents a great majority of the users; this has been enforced when dividing by age, since the range of young customers (less than 20 years old) was the most recurrent. The distribution of the two genders was quite balanced. Moreover, the temporal distribution of the journeys of retired people followed a completely different trend with respect to the other categories.
- **Travel document.** Subscriptions, especially those reserved for students and in general with a year of validity, were the most frequent category, while tickets (single or carnet) were quite rare. Instead, the situation was reversed on holidays.

- **Stop point and route.** Validations have turned out to be concentrated in the towns of Cuneo and Saluzzo, which were also the places with the highest density of stop points.
- **Trip.** In this case, the distribution of the validations was more balanced: the most recurrent trips weren't ran on holidays and many of them showed a decline of the demand on Saturdays.
- **Weather.** With the premise that historical weather data were not very detailed, especially in time, thus making difficult to compare the demand across different weather scenarios, some of which were very rare or completely absent during the whole month, it seems that bad conditions enhanced people to use public transport.

The focus has then been moved on the single stop points, split by route. They have been clustered, separately for each type of day, based on the hourly number of incoming validations and on its variance. Then, variables related to the features of each stop and to the categories of customers validating at each of them have been added. This has allowed to observe that the majority of the samples formed a very compact group, where the importance of the stops was quite low. Then, it has been interesting to find out the differences among the other groups: it has emerged that the most characterising variables were the number of trips passing in average by that stop point and the demand coming from students and *others*. Moreover, from the geographical distribution of each group, it has been clear which stop points are located in an isolated area and which are in a central position in a certain town, thus finding a certain correspondence with the importance of the stop point itself.

8.2 Possible future evolution

The two analysis conducted in this thesis should be both interpreted as a preparatory step for a definitely more important matter: the forecast of the demand at each stop point (or also, for a group of close stop points). Actually, in Chapter 5 the relationship between this target and the possible predictors has been built and in Chapter 7.1 it has been deeply observed: it can be concluded that each of them has its own importance, even if not always the same. However, as already said, there are also other variables that it is worth considering, such as the traffic conditions, the capacity of the vehicle and the occasional events, which in this context were not available.

Then, the segmentation of the stop points should be seen as a tool for distinguishing those with a lower and more stable demand, where also a very basic model is likely to perform well in the prediction, from the most crowded ones, where the demand considerably changes during the day, therefore needing more accurate and

sophisticated techniques, such as those mentioned in Chapter 2 because used in other documents which dealt with a similar problem.

Of course, some of the formed groups may have been further partitioned or joined together, depending on the desired level of similarity or on a minimum number of samples needed to form a group. Moreover, this forecast should be in turn properly exploited in order to make the service more efficient: in terms of time because it would allow to find out when it should be intensified (for example, because the bus is likely to be overcrowded); in terms of space, because it should be clear which are the most important routes, but also the most recurrent origins and destinations. Therefore, the prediction may suggest not just to modify the headway of a certain route, but also to reorganise its stops.

Bibliography

[1] Sun, Lijun & Axhausen, Kay. (2016). Understanding urban mobility patterns with a probabilistic tensor factorization framework. Transportation Research Part B Methodological. 91. 511-524. 10.1016/j.trb.2016.06.011.

[2] Tang, Jinjun & Wang, Xiaolu & Zong, Fang & Hu, Zheng. (2020). Uncovering Spatio-temporal Travel Patterns Using a Tensor-based Model from Metro Smart Card Data in Shenzhen, China. Sustainability. 12. 1475. 10.3390/su12041475.

[3] Zhou, Chunjie & Dai, Pengfei & Wang, Fusheng & Zhang, Zhenxing. (2015). Predicting the passenger demand on bus services for mobile users. Pervasive and Mobile Computing. 25. 10.1016/j.pmcj.2015.10.003.

[4] Daraio, Elena & Cagliero, Luca & Chiusano, Silvia & Garza, Paolo & Giordano, Danilo. (2020). Predicting Car Availability in Free Floating Car Sharing Systems: Leveraging Machine Learning in Challenging Contexts. Electronics. 9. 1322. 10.3390/electronics9081322.

[5] Chen, Enhui & Zhirui, Ye & Wang, Chao & Xu, Mingtao. (2019). Subway Passenger Flow Prediction for Special Events Using Smart Card Data. IEEE Transactions on Intelligent Transportation Systems. PP. 1-12. 10.1109/TITS.2019.2902405.

[6] Toqué, Florian & El Mahrsi, Mohamed & Côme, Etienne & Oukhellou, Latifa. (2016). Forecasting Dynamic Public Transport Origin-Destination Matrices with Long-Short Term Memory Recurrent Neural Networks. 10.1109/ITSC.2016.7795689.

[7] Menon, Aditya & Lee, Young. (2017). Predicting Short-Term Public Transport Demand via Inhomogeneous Poisson Processes. 2207-2210. 10.1145/3132847.3133058.

[8] Yu, Haiyang & Chen, Dongwei & Wu, Zhihai & Ma, Xiaolei & Wang, Yunpeng. (2016). Headway-based bus bunching prediction using transit smart card data. Transportation Research Part C: Emerging Technologies. 72. 10.1016/j.trc.2016.09.007.

[9] Liu, Wei & Shoji, Yozo. (2019). DeepVM: RNN-based Vehicle Mobility Prediction to Support Intelligent Vehicle Applications. IEEE Transactions on Industrial Informatics. PP. 1-1. 10.1109/TII.2019.2936507.

[10] Costa, Vera & Dias, Teresa & Costa, Pedro & Fontes, Tânia. (2015). Prediction of Journey Destination in Urban Public Transport. 10.1007/978-3-319-23485-4_18.

[11] Yu, Haiyang & Wu, Zhihai & Chen, Dongwei & Ma, Xiaolei. (2016). Probabilistic Prediction of Bus Headway Using Relevance Vector Machine Regression. IEEE Transactions on Intelligent Transportation Systems. 18. 1-10. 10.1109/TITS.2016.2620483.

[12] Qi, Geqi & Huang, Ailing & Wei, Guan & Fan, Lingling. (2018). Analysis and Prediction of Regional Mobility Patterns of Bus Travellers Using Smart Card Data and Points of Interest Data. IEEE Transactions on Intelligent Transportation Systems. PP. 1-18. 10.1109/TITS.2018.2840122.

[13] Song, Xuan, H. Kanasugi and R. Shibasaki. "DeepTransport: Prediction and Simulation of Human Mobility and Transportation Mode at a Citywide Level." IJCAI (2016).

[14] Sun, Fangzhou & Pan, Yao & White, Jules & Dubey, Abhishek. (2016). Real-Time and Predictive Analytics for Smart Public Transportation Decision Support System. 10.1109/SMARTCOMP.2016.7501714.

[15] Zhao, Zhan & Koutsopoulos, Haris & Zhao, Jinhua. (2018). Individual mobility prediction using transit smart card data. Transportation Research Part C Emerging Technologies. 89. 19-34. 10.1016/j.trc.2018.01.022.

[16] Othman, Md. Shalihin & Tan, Gary. (2018). Predictive Simulation of Public Transportation Using Deep Learning: 18th Asia Simulation Conference, AsiaSim 2018, Kyoto, Japan, October 27–29, 2018, Proceedings. 10.1007/978-981-13-2853-4_8.

[17] Cristóbal, Teresa & Padrón, Gabino & Quesada-Arencibia, Alexis & Hernández, Francisco & Blasio, Gabriel & García, Carmelo. (2018). Using Data Mining to Analyze Dwell Time and Nonstop Running Time in Road-Based Mass Transit Systems. Proceedings. 2. 1217. 10.3390/proceedings2191217.

[18] Li, Junfang & Yao, Minfeng & Fu, Qian. (2016). Forecasting Method for Urban Rail Transit Ridership at Station Level Using Back Propagation Neural Network. Discrete Dynamics in Nature and Society. 2016. 1-9. 10.1155/2016/9527584.

[19] Carpio-Pinedo, Jose. (2014). Urban Bus Demand Forecast at Stop Level: Space Syntax and Other Built Environment Factors. Evidence from Madrid. Procedia - Social and Behavioral Sciences. 160. 10.1016/j.sbspro.2014.12.132.

[20] Gutiérrez, Javier & Cardozo, Osvaldo Daniel & García-Palomares, Juan. (2011). Transit ridership forecasting at station level: An approach based on distance-decay weighted regression. Journal of Transport Geography. 19. 1081-1092. 10.1016/j.jtrangeo.2011.05.004.

[21] Asmael, Noor & Waheed, Mohanned. (2018). Demand estimation of bus as a public transport based on gravity model. MATEC Web of Conferences. 162.

01038. 10.1051/matecconf/201816201038.

[22] Palacio, Sebastián. (2018). Machine Learning Forecasts of Public Transport Demand: A Comparative Analysis of Supervised Algorithms Using Smart Card Data. SSRN Electronic Journal. 10.2139/ssrn.3165303.

[23] Eleftherios Kofidis. Tensor Methods for Multi-Aspect Trajectory Data Mining. 2020. hal-02639989

[24] Kolda, Tamara & Bader, Brett. (2009). Tensor Decompositions and Applications. SIAM Review. 51. 455-500. 10.1137/07070111X.

[25] Briand, Anne-Sarah & Côme, Etienne & Trépanier, Martin & Oukhellou, Latifa. (2017). Analyzing year-to-year changes in public transport passenger behaviour using smart card data. Transportation Research Part C: Emerging Technologies. 79. 274-289. 10.1016/j.trc.2017.03.021.

[26] Arnone, Maurizio & Delmastro, Tiziana & Giacosa, Giulia & Paoletti, Mauro & Villata, Paolo. (2016). The Potential of E-ticketing for Public Transport Planning: The Piedmont Region Case Study. Transportation Research Procedia. 18. 3-10. 10.1016/j.trpro.2016.12.001.

[27] Arneodo, Fabrizio & Arnone, Maurizio & Botta, Danilo & Delmastro, Tiziana & Negrino, Carola. Estimation of public transport user behaviour and trip chains through the Piedmont Region e-ticketing system, Paper number ITS-SP2273

[28] Li, T. & Sun, D. & Jing, P. & Yang, K. Smart Card Data Mining of Public Transport Destination: A Literature Review. Information 2018, 9, 18

[29] M. Trépanier & N. Tranchant & R. Chapleau. (2007). Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System. Journal of Intelligent Transportation Systems: Technology, Planning, and Operations 11 (1), pp. 1-14.

[30] L. He & M. Trépanier (2015). Estimating the destination of unlinked trips in transit smart card fare data. Transportation Research Record: Journal of the Transportation Research Board 2535, pp. 97-104

[31] J. Jung & K. Sohn (2017). Deep Learning Architecture to Forecast Destinations of Bus Passengers from Entry only Smart card Data. IET Intelligent Transport Systems , Vol. 11, pp. 334-339.

[32] Falchetta, Giacomo & Noussan, Michel. (2020). The Impact of COVID-19 on Transport Demand, Modal Choices, and Sectoral Energy Consumption in Europe.

[33] Short, Eleanor & Gouge, Taylor & Mills, Gareth (2020). Public transport and Covid-19, How to transition from response to recovery

[34] https://www.3bmeteo.com

[35] http://www.comuni-italiani.it

[36] https://opencagedata.com/

[37] https://mrcagney.github.io/gtfs_kit_docs/index.html#module-gtfs_kit.stops
[38] https://www.istat.it/it/archivio/104317#accordions
[39] https://www.istat.it/storage/codici-unita-amministrative/Elenco-comuni-italiani.xls
[40] https://www.istat.it/it/files//2011/01/Elenco-codici-e-denominazioni-unita-territoriali-estere.zip
[41] http://download.geofabrik.de/europe/italy.html